A SYSTEM APPROACH TO MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION AND RESIDUAL ECHO SUPPRESSION FOR ROBUST HANDS-FREE TELECONFERENCING

A Thesis
Presented to
The Academic Faculty

by

Jason Wung

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2015

A SYSTEM APPROACH TO MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION AND RESIDUAL ECHO SUPPRESSION FOR ROBUST HANDS-FREE TELECONFERENCING

Approved by:

Professor Biing-Hwang (Fred) Juang,
Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Gordon L. Stüber
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Mark A. Clements
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Professor Brani Vidakovic
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor David V. Anderson
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Date Approved: March 3, 2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Motivations

Acoustic echo aries due to the coupling between loudspeakers and microphones in the same room during hands-free teleconferencing. The sound that is played from the loudspeakers in the near-end room will be convolved with the near-end room impulse responses (RIRs), picked up by the microphones in the same room, and eventually transmitted back to the far-end room. Due to the round-trip transmission delay, this acoustic echo is very distracting and severely limits the quality of the conversation. An acoustic echo reduction system is required in such situations to ensure high quality hands-free teleconferencing.

In such scenarios, an adaptive filter is a powerful tool to keep track of the time variations of the unknown system, i.e., the near-end RIRs, and to produce an *estimate* of the acoustic echo at the near-end room in attempt to cancel the echo. However, due to the mismatch between the true and the estimated near-end RIRs, acoustic echo cancellation (AEC) alone is not sufficient and residual echo suppression (RES) is often required to further suppress the echo that cannot be entirely canceled by the AEC.

Figure 1 shows a single-channel acoustic echo reduction system with an adaptive filter $\mathbf{w}$ for AEC and a postfilter H for RES. Let y be the near-end microphone signal, which consists of the near-end noise or speech v mixed with the acoustic echo $d = \mathbf{h}^{\mathrm{T}}\mathbf{x}$, where $\mathbf{h}$ is the RIR vector (a truncated version of the actual RIR),[1] $\mathbf{x}$ is

[1] Although in reality the RIR can be infinite in length, for simplification of the analysis of adaptive filter algorithms $\mathbf{h}$ is assumed to be of finite length L unless otherwise specified.

the far-end reference signal vector, and $\{\cdot\}^{\mathrm{T}}$ is the transpose operator.

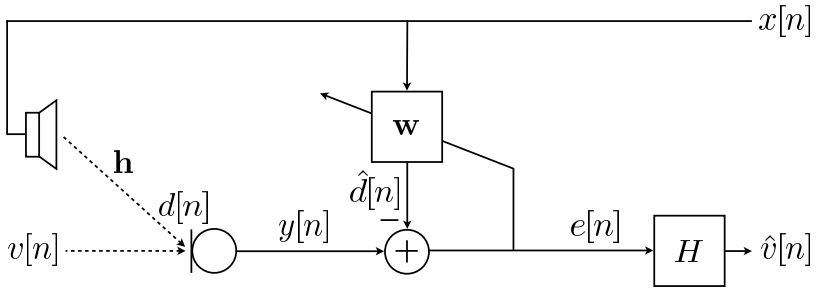


Figure 1: An acoustic echo reduction system with an adaptive filter $\mathbf{w}$ for AEC and a postfilter H for RES.

The adaptive filter coefficients $\mathbf{w}$ model the RIR, and the filtered output $\hat{d} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ approximates the echo d. The observed estimation error e of the AEC is given by

$$e[n] = v[n] + d[n] - \hat{d}[n] = v[n] + b[n], \tag{1}$$

where n is the time-domain sample index and b is the *true* error (residual echo) that comes from the misalignment between the RIR and the estimated adaptive filter coefficients, i.e., $(\mathbf{h} - \mathbf{w})^{\mathrm{T}}\mathbf{x}$. The term "true" means a noise-free quantity, i.e., $v = 0$. A postfilter H further suppresses the residual echo before the processed near-end signal is sent to the far-end room. The presence of a strong near-end signal v may disrupt the tracking of the near-end RIR, resulting in a corrupted error signal b. Traditionally, a double-talk detector (DTD) is used to detect such a signal mixing environment, i.e., to determine if the near-end signal is present, and once so detected, to freeze the adaptation of the filter coefficients.

The acoustic echo reduction system also finds applications in other application scenarios such as multiplayer online gaming or voice commands for smart televisions (TVs). In multiplayer online gaming, the system (game consoles) actively produces music or sound effects that may interfere with the communication among players at different sites. Specifically, both the music or the sound effects from the game and the voice of the user in the near-end room will be captured by the microphones and

transmitted to the far-end room. The acoustic "echo" in this case consists of the music and sound effects from the game. For smart TVs, a user may issue voice commands to control the TV at any given moment without pausing the program that is currently played on the set. In such a scenario, the audio from the TV program constitutes the acoustic "echo" in the eyes of the voice command unit. In both cases, the microphone signals, which will be either transmitted to the far-end for communication purposes or recognized by the local machines, are contaminated by the sound produced by the system itself and an effective acoustic echo reduction system can either enhance the speech quality for communication or increase the speech recognition performance.

In these scenarios, the signal of interest, v, always causes the system to operate in a double-talk mode since the loudspeaker will continuously play back either music and sound effects from games or audio from TV programs. The traditional usage of a DTD to freeze the adaptive filter during double talk no longer provides adequate echo cancellation for such scenarios. Therefore, noise robust adaptive algorithms that allow the adaptive filter coefficients to be updated continuously in the presence of a strong near-end signal are very desirable. In addition, these systems often employ multiple loudspeakers and microphones and thus the echo reduction system must perform multi-channel acoustic echo cancellation (MCAEC) with multiple adaptive filters, each of which models the echo path from each loudspeaker to each microphone. Due to the high correlation nature of the loudspeaker signals, the so called *non-uniqueness problem* causes the convergence speed to slow down significantly for MCAEC.

1.2 Objectives

The objective of the research is to achieve a systematic combination of acoustic echo reduction components that together achieve a robust performance of the MCAEC system as a whole. Conventional approaches to the acoustic echo reduction system typically assume that individual components would perform ideally. For example, the

adaptive algorithm for AEC is often developed in the absence of strong near-end signal, the algorithm for RES is often an added module that is developed as a separate noise reduction component, and the decorrelation procedure for MCAEC is yet another add-on module that simply introduces some form of distortion to the reference signal. Specifically, the signal of interest in the AEC component is the error signal (i.e., the signal after the subtraction of the estimated echo), proper calculation of which is marred by the presence of the near-end signal which acts as the interference, or noise, to the adaptive filter in the AEC component. On the other hand, the signal of interest of the RES component is the near-end signal which needs to be preserved for transmission to the far-end while the residual echo is being suppressed. The main challenge is in designing a consistent criterion across all modules that can be jointly optimized to form a more consistent framework for acoustic echo reduction. The decorrelation procedure can potentially benefit from the system approach as well if it is designed by taking the near-end listener into account. The MCAEC system should be optimized not only for the echo cancellation and suppression performance, but also for the reference signal quality after the added distortion from the decorrelation procedure.

1.3 Outline

The rest of this dissertation is organized as follows.

In Chapter 2, we review the least mean squares (LMS) and the normalized least mean squares (NLMS) algorithms that is traditionally used in AEC. The frequency-domain processing framework for the AEC is also introduced. We then discuss the non-uniqueness problem and the misalignment problem that occur in MCAEC. We review algorithms that deal with the double talk situation and related robustness issues. We then provide the single-channel speech enhancement framework that deals with the nonlinear processing of residual echo and noise suppression after the AEC.

We also define the performance measures for AEC and RES and the coherence measure for MCAEC.

In Chapter 3, we present the system approach to AEC and RES for the single-channel case. To include RES in the system approach, we propose a novel estimation procedure for the residual echo that utilizes both the linearly estimated echo signal from the adaptive filter and the nonlinearly estimated echo from the noise suppression (NS) unit. We then propose a novel combination of the AEC and the RES to form an system approach to AEC that achieves effective echo reduction even during double talk. We finally show the application of the system approach to AEC on the *Kinect*™ audio and demonstrate the superior performance of the voice quality enhancement system both in terms of objective quality measures and automatic speech recognition results.

In Chapter 4, we present a novel sub-band resampling (SBR) approach suitable for the decorrelation procedure in MCAEC. We first formulate the misalignment problem and show how the signal coherence affects the misalignment. We then propose a novel resampling procedure and perform a deep analysis to show the link between resampling, coherence, and misalignment. We discuss the algorithmic design issues and proper smoothing method to avoid possible processing artifacts from the resampling procedure. We then extend the resampling technique to SBR so that the coherence can be finely adjusted in each frequency bin. In doing so the decorrelated speech signal can retain as high fidelity as possible while significantly improving the convergence rate of the MCAEC. We demonstrate the superior sound quality of SBR through objective quality measures as well as a formal listening test.

In Chapter 5, instead of looking at individual components, i.e., AEC, RES, or decorrelation, we piece together all the components to build a robust multi-channel voice system that aims at achieving the best voice quality with a given computational complexity constraint, suitable for real-time applications on many different platforms.

We first discuss a complete signal-channel voice system that includes the components designed in Chapter 3. We then formulate the tuning of the signal-channel voice system as a constrained optimization problem to optimize the system for best voice quality within the computational budget of a target platform. Finally, we include the SBR approach described in Chapter 4 and present a robust regularization procedure for MCAEC.

In Chapter 6, we conclude by summarizing our contributions and discuss directions for future works.

CHAPTER II

PROBLEM BACKGROUND

This chapter is organized as follows. In Section 2.1, we review the algorithms that are suitable for the acoustic echo cancellation (AEC) problem. Specifically, we review the least mean squares (LMS) algorithm, the normalized least mean squares (NLMS) algorithm, and their frequency-domain variants. We then review the multi-channel acoustic echo cancellation (MCAEC) problem in Section 2.2 and formulate the non-uniqueness and the misalignment problem. In Section 2.3, we review the robust acoustic echo cancellation (RAEC) and its various components that are suitable for our applications, i.e., to enhance the speech signal even during continuous double talk situation. In Section 2.4, we review the single-channel noise reduction framework that is necessary to further suppress the residual echo, which often results from the mismatch of the estimated impulse response in the AEC. In Section 2.5, we introduce several performance measures that are suitable for the evaluation of the AEC, the residual echo and noise suppressor, and the coherence measure that is useful for measuring the correlation between the loudspeaker signals in MCAEC.

2.1 Acoustic Echo Cancellation

Assuming that the additive noise v is zero in (1), the error signal is expressed as

$$e[n] = b[n] = d[n] - \mathbf{w}^{\mathrm{T}}[n]\mathbf{x}[n], \tag{2}$$

where d is the desired signal (acoustic echo), $\mathbf{w}[n] = [w_0[n], \ldots, w_{L-1}[n]]^{\mathrm{T}}$ is the adaptive filter coefficients vector of length L, and $\mathbf{x}[n] = [x[n], \ldots, x[n-L+1]]^{\mathrm{T}}$ is the reference signal vector.

Gradient descent is an optimization algorithm that finds a local minimum of a

cost function by taking steps proportional to the *negative* of the gradient of the cost function at the current point, i.e.,

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \mu \nabla_{\mathbf{w}^*} J(\mathbf{w}[n]), \tag{3}$$

where μ is the step-size, $\{\,\cdot\,\}^*$ is the element-wise complex conjugate operator, and $\nabla_{\mathbf{w}^*}$ is the gradient operator with respect to $\mathbf{w}^*$ defined as $\nabla_{\mathbf{w}^*} J \equiv \left[\frac{\partial J}{\partial w_0^*}, \ldots, \frac{\partial J}{\partial w_{L-1}^*}\right]^{\mathrm{T}}$ with $J(\,\cdot\,)$ being the cost function.[1]

Based on the choice of the cost function, there are basically three types of gradient-descent based adaptive algorithms, i.e., the LMS, the affine projection (AP), and the recursive least squares (RLS) algorithms [99, 100]. Table 1 shows a comparison of the three gradient-descent based adaptive algorithms, where $\mathrm{E}\{\,\cdot\,\}$ is the expectation operator, L is the length of the adaptive filter, Q is the AP order, and $0 \ll \lambda < 1$ is the "forgetting factor" which gives exponentially less weight to older error samples. Note that for the RLS algorithm, the input signals are considered deterministic, while for the LMS and the AP algorithms they are considered stochastic. Compared to the LMS and the AP algorithms, the RLS algorithm exhibits extremely fast convergence speed at the expense of high computational complexity and numerical instability. On the other hand, the LMS algorithm is widely used for the AEC and will be the focus of this work due to its low computational complexity, numerical stability, and ease of implementation.

Table 1: A comparison of three gradient-descent based adaptive algorithms that are commonly used for AEC: the LMS, the AP, and the RLS algorithms.

Algorithm	Cost Function	Complexity
LMS	$\mathrm{E}\{\lvert e[n]\rvert^2\}$	$\mathcal{O}\{L\}$
AP	$\mathrm{E}\{\sum_{l=0}^{Q-1} \lvert e[n-l]\rvert^2\}$	$\mathcal{O}\{LQ\}$
RLS	$\sum_{l=0}^{n} \lambda^{n-l} \lvert e[l]\rvert^2$	$\mathcal{O}\{L^2\}$

Faster convergence and computational efficiency of the LMS algorithm can be

[1] For a complex vector $\mathbf{w}$, the partial gradient with respect to $\mathbf{w}^*$ gives the direction of maximum rate of change [13], rather than the gradient with respect to $\mathbf{w}$.

achieved by using block-based frequency-domain least mean squares (FDLMS) algorithms [102] instead of sample-based time-domain LMS algorithm. Faster and more uniform convergence is achieved through the diagonalization in the discrete Fourier transform (DFT) domain so that the adaptation step-size can be adjusted independently for each frequency bin, while the computational efficiency is achieved through the efficient implementation of the convolution operation with DFT via the overlap-add or the overlap-save (OLS) method [92].

2.1.1 Least Mean Squares

For the analysis of the LMS-type algorithms, all signals are assumed to be generated by zero-mean random processes, and the near-end noise v is assumed to be zero. By omitting the sample index for simplicity, the cost function using the mean squared error (MSE) can be expanded as

$$\begin{aligned} J(\mathbf{w}) &= \mathrm{E}\{|e|^2\} = \mathrm{E}\{(d - \mathbf{w}^{\mathrm{T}}\mathbf{x})(d - \mathbf{w}^{\mathrm{T}}\mathbf{x})^*\} \\ &= \sigma_d^2 - \mathbf{r}_{d\mathbf{x}}^{\mathrm{H}}\mathbf{w} - \mathbf{w}^{\mathrm{H}}\mathbf{r}_{d\mathbf{x}} + \mathbf{w}^{\mathrm{H}}\mathbf{R}_{\mathbf{x}}\mathbf{w}, \end{aligned} \tag{4}$$

where $\{\,\cdot\,\}^{\mathrm{H}}$ is the Hermitian transpose (conjugate transpose) operator, $\sigma_d^2 \equiv \mathrm{E}\{dd^*\}$ is the variance of the signal d, $\mathbf{r}_{d\mathbf{x}} \equiv \mathrm{E}\{d\mathbf{x}^*\}$ is the cross-covariance vector between d and $\mathbf{x}$, and $\mathbf{R}_{\mathbf{x}} \equiv \mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathrm{H}}\}$ is the autocorrelation matrix of $\mathbf{x}$. The gradient of the cost function is given by

$$\nabla_{\mathbf{w}^*} J(\mathbf{w}[n]) = -(\mathbf{r}_{d\mathbf{x}} - \mathbf{R}_{\mathbf{x}}\mathbf{w}[n]). \tag{5}$$

Using (3) and (5), the update rule is given by

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu(\mathbf{r}_{d\mathbf{x}} - \mathbf{R}_{\mathbf{x}}\mathbf{w}[n]). \tag{6}$$

Note that (6) requires the statistical information of $\mathbf{r}_{d\mathbf{x}}$ and $\mathbf{R}_{\mathbf{x}}$ which are rarely available in practice. As a result, the instantaneous values $\mathbf{r}_{d\mathbf{x}} \approx d\mathbf{x}^*$ and $\mathbf{R}_{\mathbf{x}} \approx \mathbf{x}\mathbf{x}^{\mathrm{H}}$

are used and the LMS algorithm is given by

$$\begin{aligned}\mathbf{w}[n+1] &= \mathbf{w}[n] + \mu(d[n]\mathbf{x}^*[n] - \mathbf{x}[n]\mathbf{x}^{\mathrm{H}}[n]\mathbf{w}[n]) \\ &= \mathbf{w}[n] + \mu(d[n] - \mathbf{w}^{\mathrm{T}}[n]\mathbf{x}[n])\mathbf{x}^*[n] \\ &= \mathbf{w}[n] + \mu e[n]\mathbf{x}^*[n]. \end{aligned} \tag{7}$$

The LMS algorithm is summarized in Table 2, where $\mathbf{0}_{L\times 1} \equiv [0, \ldots, 0]^{\mathrm{T}}$ is a zero vector of length L.

Table 2: The LMS algorithm.

Initialization
$\mathbf{w}[0] = \mathbf{0}_{L\times 1}$
$\{x[n] = 0; n < 0\}, \quad L > 0, \quad \mu > 0$
Filter adaptation
$\mathbf{x}[n] = [x[n], \ldots, x[n-L+1]]^{\mathrm{T}}$
$e[n] = d[n] - \mathbf{w}^{\mathrm{T}}[n]\mathbf{x}[n]$
$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu e[n]\mathbf{x}^*[n]$

2.1.2 Newton's Method

The gradient update of the filter coefficients in (3) can actually be of the more general form

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \mu\mathbf{u}[n], \tag{8}$$

where $\mathbf{u} = \mathbf{B}\nabla_{\mathbf{w}^*}J(\mathbf{w})$ and $\mathbf{B}$ is any Hermitian positive-definite matrix. Using (4), (5), and (8), the MSE after each iteration can be written as

$$\begin{aligned} J(\mathbf{w}[n+1]) &= \sigma_d^2 - \mathbf{r}_{d\mathbf{x}}^{\mathrm{H}}\mathbf{w}[n+1] - \mathbf{w}^{\mathrm{H}}[n+1]\mathbf{r}_{d\mathbf{x}} + \mathbf{w}^{\mathrm{H}}[n+1]\mathbf{R}_{\mathbf{x}}\mathbf{w}[n+1] \\ &= J(\mathbf{w}[n]) - \mu\Re\{\mathbf{u}^{\mathrm{H}}[n]\nabla_{\mathbf{w}^*}J(\mathbf{w}[n])\} + \mu^2\mathbf{u}^{\mathrm{H}}[n]\mathbf{R}_{\mathbf{x}}\mathbf{u}[n], \end{aligned} \tag{9}$$

where $\Re\{\cdot\}$ is the real value of a complex number. Note that with $\mathbf{B}$ being Hermitian positive-definite, the middle term in (9) is strictly negative and

$$J(\mathbf{w}[n+1]) < J(\mathbf{w}[n]). \tag{10}$$

The previous derivation of the LMS algorithm can be viewed as a special case of using the identity matrix $\mathbf{B} = \mathbf{I}_{L\times L} \equiv \text{diag}\{\mathbf{1}_{L\times 1}\}$, where $\text{diag}\{\cdot\}$ is an operator that forms a diagonal matrix and $\mathbf{1}_{L\times 1} \equiv [1, \ldots, 1]^{\text{T}}$.

Another useful choice of the matrix $\mathbf{B}$ is the inverse of the complex Hessian matrix $\mathbf{B} = \left[\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}^* \partial \mathbf{w}^{\text{T}}}\right]^{-1}$, which leads to the Newton's method

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \mu \left[\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}^* \partial \mathbf{w}^{\text{T}}}\right]^{-1} \nabla_{\mathbf{w}^*} J(\mathbf{w}[n]). \tag{11}$$

Using (5), (11), and the MSE, the Newton's method becomes

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu(\epsilon \mathbf{I}_{L\times L} + \mathbf{R}_{\mathbf{x}})^{-1}(\mathbf{r}_{d\mathbf{x}} - \mathbf{R}_{\mathbf{x}}\mathbf{w}[n]), \tag{12}$$

where $\epsilon > 0$ is a regularization term for stability when the Hessian matrix is close to singular. Note that by choosing $\epsilon = 0$ and $\mu = 1$, (12) leads to immediate convergence since

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{r}_{d\mathbf{x}} - \mathbf{w}[n] = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{r}_{d\mathbf{x}}, \tag{13}$$

which is the optimum least squares solution to the MSE that can be obtained by setting the gradient (5) to zero and solve for $\mathbf{w}$. This is a well-known property of the Newton's method where the convergence is guaranteed in a single iteration by choosing $\mu = 1$ [100].

2.1.3 Normalized Least Mean Squares

By using the instantaneous values $\mathbf{R}_{d\mathbf{x}} \approx d\mathbf{x}^*$ for the cross-correlation vector and $\mathbf{R}_{\mathbf{x}} \approx \mathbf{x}\mathbf{x}^{\text{H}}$ for the autocorrelation matrix in (12), the update rule can be obtained similarly to the LMS algorithm as

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu e[n](\epsilon \mathbf{I}_{L\times L} + \mathbf{x}[n]\mathbf{x}^{\text{H}}[n])^{-1}\mathbf{x}^*[n], \tag{14}$$

which requires the inversion of the matrix $\epsilon \mathbf{I}_{L\times L} + \mathbf{x}\mathbf{x}^{\mathrm{H}}$ at each iteration. By using the matrix inversion formula, the inversion can be simplified as

$$(\epsilon \mathbf{I}_{L\times L} + \mathbf{x}[n]\mathbf{x}^{\mathrm{H}}[n])^{-1} = \epsilon^{-1}\mathbf{I}_{L\times L} - \frac{\epsilon^{-2}}{1+\epsilon^{-1}\|\mathbf{x}[n]\|^2}\mathbf{x}[n]\mathbf{x}^{\mathrm{H}}[n], \tag{15}$$

where $\|\cdot\|$ is the Euclidean norm operator defined as $\|\mathbf{x}\| \equiv \sqrt{\mathbf{x}^{\mathrm{H}}\mathbf{x}}$. Therefore, the NLMS algorithm with regularization is given by

$$\begin{aligned}\mathbf{w}[n+1] &= \mathbf{w}[n] + \mu e[n]\left(\epsilon^{-1}\mathbf{x}^*[n] - \frac{\epsilon^{-2}}{1+\epsilon^{-1}\|\mathbf{x}[n]\|^2}\mathbf{x}^*[n]\|\mathbf{x}[n]\|^2\right) \\ &= \mathbf{w}[n] + \mu e[n]\frac{\mathbf{x}^*[n]}{\epsilon + \|\mathbf{x}[n]\|^2}.\end{aligned} \tag{16}$$

Table 3 shows a summary of the NLMS algorithm with regularization.

Table 3: The NLMS algorithm with regularization.

Initialization

$$\begin{aligned}&\mathbf{w}[0] = \mathbf{0}_{L\times 1} \\ &\{x[n] = 0; n < 0\}, \quad L > 0, \quad \mu > 0, \quad \epsilon > 0\end{aligned}$$

Filter adaptation

$$\begin{aligned}\mathbf{x}[n] &= [x[n], \ldots, x[n-L+1]]^{\mathrm{T}} \\ e[n] &= d[n] - \mathbf{w}^{\mathrm{T}}[n]\mathbf{x}[n] \\ \mathbf{w}[n+1] &= \mathbf{w}[n] + \mu e[n]\frac{\mathbf{x}^*[n]}{\epsilon + \|\mathbf{x}[n]\|^2}\end{aligned}$$

The NLMS algorithm is more robust to the LMS algorithm in that the update of the filter coefficients is invariant to the scale of the reference signal vector $\mathbf{x}$, since in the LMS algorithm the update of the filter coefficients $\mathbf{w}$ will be proportional to the norm of the reference signal $\mathbf{x}$ and the update between iterations can fluctuate significantly depending on how large or small the norm is. Such a behavior can have an adverse effect on the performance of the LMS algorithm when dealing with speech signals as the reference signal, since the norm of the speech signals varies widely between speech activity and silence. Besides being more stable than the LMS

algorithm, the NLMS algorithm also exhibits faster convergence due to the superior convergence speed of the Newton's method. The NLMS algorithm can be seen as employing a time-varying step-size $\mu/(\epsilon+\|\mathbf{x}[n]\|^2)$ as opposed to the constant step-size μ in the LMS algorithm. The adjustment of the step-size is indeed very important for the robust operation of the adaptive filter and will be further discussed in Section 2.3.

2.1.4 Frequency-Domain Least Mean Squares

The frequency-domain version of the LMS algorithm using the OLS procedure can in general be written as

$$\underline{\mathbf{w}}[m+1] = \underline{\mathbf{w}}[m] + \mu\mathbf{G}^{10}(\underline{\mathbf{e}}[m] \circ \underline{\mathbf{x}}^*[m]), \tag{17}$$

where m is the frame index, $\underline{\mathbf{w}}[m] = \mathbf{F}[\mathbf{w}^{\mathrm{T}}[m], \mathbf{0}_{L\times 1}^{\mathrm{T}}]^{\mathrm{T}}$ is the zero-padded filter coefficient vector transformed to the frequency domain,[2] $\mathbf{F}$ is a $2L \times 2L$ DFT matrix with elements $[\mathbf{F}]_{k+1,n+1} \equiv \exp(-j\frac{2\pi kn}{2L}) = \omega_{2L}^{kn}$ in the $(k+1)^{\text{th}}$ row and the $(n+1)^{\text{th}}$ column, $\omega_{2L} \equiv \exp(-j\frac{2\pi}{2L})$, $\exp(\,\cdot\,)$ is the exponential function, $\mathbf{G}^{10}$ is a $2L \times 2L$ gradient constraint matrix, $\underline{\mathbf{x}}[m] = \mathbf{F}[\mathbf{x}^{\mathrm{T}}[m-1], \mathbf{x}^{\mathrm{T}}[m]]^{\mathrm{T}}$ is the DFT of two reference signal blocks, $\underline{\mathbf{e}}[m] = \mathbf{F}[\mathbf{0}_{L\times 1}^{\mathrm{T}}, \mathbf{e}^{\mathrm{T}}[m]]^{\mathrm{T}}$ is the DFT of the zero-padded error signal block, and $\circ$ is the Hadamard product (element-wise multiplication) defined as $[\mathbf{A}\circ\mathbf{B}]_{i,j} \equiv A_{i,j}B_{i,j}$ in the i^{th} row and the j^{th} column for all $M \times N$ matrices $\mathbf{A}$ and $\mathbf{B}$. Note that the reference signal and the error signal blocks of the m^{th} frame are given by $\mathbf{x}[m] = [x[mL], \ldots, x[(m+1)L-1]]^{\mathrm{T}}$ and $\mathbf{e}[m] = [e[mL], \ldots, e[(m+1)L-1]]^{\mathrm{T}}$, respectively.

The zero-padding is required for the proper filtering operation in the frequency domain by using the convolution theorem. Since the Hadamard product of two DFT coefficient vectors result in a circular convolution in the time domain, the zero-padding and the gradient constraint matrix together remove the aliasing artifacts from the circular convolution. This is achieved by properly sectioning the signals in the time

[2]Note that an underlined vector denotes a frequency domain quantity.

domain and then transform the sectioned signal back to the frequency domain. By using the circular convolution, filtering of $\mathbf{x}$ by $\mathbf{w}$ can be expressed in terms of matrix-vector multiplication as

$$\begin{bmatrix} \mathbf{a}[m] \\ \hat{\mathbf{d}}[m] \end{bmatrix} = \mathbf{W}[m] \begin{bmatrix} \mathbf{x}[m-1] \\ \mathbf{x}[m] \end{bmatrix}, \tag{18}$$

where $\hat{\mathbf{d}}[m] = [\hat{d}[mL], \ldots, \hat{d}[(m+1)L-1]]^{\mathrm{T}}$ is the desired time-domain estimated echo vector, $\mathbf{a}$ is an undesired aliasing vector, and $\mathbf{W}$ is a $2L \times 2L$ circular convolution matrix given by (the frame index is omitted for simplicity)

$$\mathbf{W} = \begin{bmatrix} w_0 & 0 & \ldots & \ldots & 0 & w_{L-1} & \ldots & w_1 \\ \vdots & \ddots & \ddots & & & \ddots & \ddots & \vdots \\ w_{L-2} & \ldots & w_0 & 0 & \ldots & \ldots & 0 & w_{L-1} \\ w_{L-1} & \ldots & \ldots & w_0 & 0 & \ldots & \ldots & 0 \\ 0 & w_{L-1} & \ldots & \ldots & w_0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & w_{L-1} & \ldots & \ldots & w_0 & 0 \\ 0 & \ldots & \ldots & 0 & w_{L-1} & \ldots & \ldots & w_0 \end{bmatrix}. \tag{19}$$

Note that in the upper half of $\mathbf{W}$, the filter coefficients are wrapped around and contribute to the aliasing vector $\mathbf{a}$. By using the circular convolution theorem, (18) can be exactly calculated by

$$\begin{bmatrix} \mathbf{a}[m] \\ \hat{\mathbf{d}}[m] \end{bmatrix} = \mathbf{F}^{-1}(\underline{\mathbf{w}}[m] \circ \underline{\mathbf{x}}[m]), \tag{20}$$

where $\mathbf{F}^{-1} \equiv \frac{1}{2L}\mathbf{F}^{*}$ is the $2L \times 2L$ inverse discrete Fourier transform (IDFT) matrix, and undesired aliasing signal can be removed by choosing the last L terms of $\mathbf{F}^{-1}(\underline{\mathbf{w}}[m] \circ \underline{\mathbf{x}}[m])$. By using the fast Fourier transform (FFT) in place of the DFT, the circular convolution can be implemented much faster than a direct calculation in the time domain.

The gradient constraint matrix $\mathbf{G}^{10}$ in (17) is given by

$$\mathbf{G}^{10} = \mathbf{F}\mathbf{D}^{10}\mathbf{F}^{-1}, \tag{21}$$

where $\mathbf{D}^{10} = \text{diag}\{[\mathbf{1}_{L\times 1}^{\text{T}}, \mathbf{0}_{L\times 1}^{\text{T}}]^{\text{T}}\}$ is a diagonal matrix for sectioning the first half of a vector. By taking the IDFT of $(\underline{\mathbf{e}} \circ \underline{\mathbf{x}}^*)$, zeroing out the second half of the resulting vector, and taking the DFT again, the update to the filter coefficients is guaranteed to have proper zero-padding in the second half of the update in the time domain. This subsequently ensures the proper operation of the circular convolution in (20).

The FDLMS algorithm is summarized in Table 4. Although the FDLMS algorithm in general exhibits better convergence property than the time-domain counterpart, one draw back is the long delay L that is characteristic of the blocking operation. In typical AEC the filter length can be as long as several hundred milliseconds which, depending on the applications, can at times be too long.

Table 4: The FDLMS algorithm.

Definitions

$$[\mathbf{F}]_{k+1,n+1} = \exp(-j\tfrac{\pi kn}{L}), \quad k,n = 0,\ldots,2L-1$$
$$\mathbf{G}^{01} = \mathbf{F}\begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}$$

Initialization

$$\underline{\mathbf{w}}[0] = \mathbf{0}_{2L\times 1}$$
$$\{x[n] = 0; n < 0\}, \quad L > 0, \quad \mu > 0$$

Filter adaptation

$$\underline{\mathbf{x}}[m] = \mathbf{F}[x[(m-1)L],\ldots,x[(m+1)L-1]]^{\text{T}}$$
$$\underline{\mathbf{d}}[m] = \mathbf{F}[\mathbf{0}_{1\times L}, d[mL],\ldots,d[(m+1)L-1]]^{\text{T}}$$
$$\underline{\mathbf{e}}[m] = \underline{\mathbf{d}}[m] - \mathbf{G}^{01}(\underline{\mathbf{w}}[m] \circ \underline{\mathbf{x}}[m])$$
$$\underline{\mathbf{w}}[m+1] = \underline{\mathbf{w}}[m] + \mu\mathbf{G}^{10}(\underline{\mathbf{e}}[m] \circ \underline{\mathbf{x}}^*[m])$$

To minimize the delay, other variants such as the multi-delay adaptive filter

[2, 3, 105] or the generalized multi-delay adaptive filter [90] can be used. These algorithms achieve lower delay by partitioning the adaptive filter into smaller blocks, while preserving the benefit of improved computational efficiency and faster convergence rate compared to the time-domain LMS algorithm. The computational complexity of FDLMS can be even lowered with the unconstrained frequency-domain adaptive filter [83] that uses only three FFT operations per block instead of five.

2.1.5 Frequency-Domain Normalized Least Mean Squares

The frequency-domain normalized least mean squares (FDNLMS) algorithm is obtained by modifying the update rule of (17) to

$$\underline{\mathbf{w}}[m+1] = \underline{\mathbf{w}}[m] + \mu \mathbf{G}^{10}(\underline{\mathbf{n}}^{\circ(-1)}[m] \circ \underline{\mathbf{e}}[m] \circ \underline{\mathbf{x}}^{*}[m]), \tag{22}$$

where $\underline{\mathbf{n}}^{\circ(-1)}$ is a normalization vector and $\mathbf{A}^{\circ(-1)}$ is defined as the Hadamard inverse (or element-wise inverse) of a matrix $\mathbf{A}$ with each element $[\mathbf{A}^{\circ(-1)}]_{i,j} \equiv A_{i,j}^{-1}$ in the i^{th} row and the j^{th} column, if and only if $A_{i,j} \neq 0, \forall i, j$. The vector $\underline{\mathbf{n}}$ is given by

$$\underline{\mathbf{n}}[m] = \epsilon \mathbf{1}_{2L \times 1} + \underline{\mathbf{s}}_x[m], \tag{23}$$

where $\epsilon > 0$ ensures numerical stability. The estimated reference signal power $\underline{\mathbf{p}}$ is given by

$$\underline{\mathbf{s}}_x[m] = \beta \underline{\mathbf{s}}_x[m-1] + (1-\beta)(\underline{\mathbf{x}}[m] \circ \underline{\mathbf{x}}^{*}[m]), \tag{24}$$

where $0 \ll \beta < 1$ is the forgetting factor that controls the effective memory of the reference signal power estimate. The FDNLMS algorithm is summarized in Table 5.

2.2 Multi-Channel Acoustic Echo Cancellation

To simplify the analysis of the MCAEC, only the echo paths that are associated with one of the microphones are considered and analyzed. Assuming the additive noise v is zero, the microphone signal of the MCAEC can in general be expressed as

$$y[n] = d[n] = \sum_{p=1}^{P} \mathbf{h}_{p,K}^{\mathrm{T}} \mathbf{x}_{p,K}[n] = \mathbf{h}_{PK}^{\mathrm{T}} \mathbf{x}_{PK}[n], \tag{25}$$

Table 5: The FDNLMS algorithm.

Definitions

$$[\mathbf{F}]_{k+1,n+1} = \exp(-j\tfrac{\pi kn}{L}), \quad k,n = 0,\ldots,2L-1$$
$$\mathbf{G}^{01} = \mathbf{F}\begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}$$

Initialization

$$\underline{\mathbf{w}}[0] = \mathbf{0}_{2L\times 1}, \quad \{x[n] = 0; n < 0\}, \quad \{\underline{\mathbf{s}}_x[m] = \mathbf{0}_{2L\times 1}; m < 0\}$$
$$L > 0, \quad \mu > 0, \quad \epsilon > 0, \quad 0 \ll \beta < 1$$

Spectral estimation

$$\underline{\mathbf{x}}[m] = \mathbf{F}[x[(m-1)L],\ldots,x[(m+1)L-1]]^{\mathrm{T}}$$
$$\underline{\mathbf{s}}_x[m] = \beta\underline{\mathbf{s}}_x[m-1] + (1-\beta)(\underline{\mathbf{x}}[m]\circ\underline{\mathbf{x}}^*[m])$$
$$\underline{\mathbf{n}}[m] = \epsilon\mathbf{1}_{2L\times 1} + \underline{\mathbf{s}}_x[m]$$

Filter adaptation

$$\underline{\mathbf{d}}[m] = \mathbf{F}[\mathbf{0}_{1\times L}, d[mL],\ldots,d[(m+1)L-1]]^{\mathrm{T}}$$
$$\underline{\mathbf{e}}[m] = \underline{\mathbf{d}}[m] - \mathbf{G}^{01}(\underline{\mathbf{w}}[m]\circ\underline{\mathbf{x}}[m])$$
$$\underline{\mathbf{w}}[m+1] = \underline{\mathbf{w}}[m] + \mu\mathbf{G}^{10}(\underline{\mathbf{n}}^{\circ(-1)}[m]\circ\underline{\mathbf{e}}[m]\circ\underline{\mathbf{x}}^*[m])$$

where P is the number of loudspeakers, d_p is the acoustic echo generated by the p^{th} loudspeaker, $\mathbf{h}_{p,K} = [h_{p,0},\ldots,h_{p,K-1}]^{\mathrm{T}}$ is the p^{th} echo path of length K from the p^{th} loudspeaker to the microphone, and $\mathbf{x}_{p,K}[n] = [x_p[n],\ldots,x_p[n-K+1]]^{\mathrm{T}}$ is the p^{th} loudspeaker signal of length K. The concatenated room impulse responses (RIRs) vector and the concatenated reference signals vector are now given as $\mathbf{h}_{PK} = [\mathbf{h}_{1,K}^{\mathrm{T}},\ldots,\mathbf{h}_{P,K}^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{x}_{PK} = [\mathbf{x}_{1,K}^{\mathrm{T}},\ldots,\mathbf{x}_{P,K}^{\mathrm{T}}]^{\mathrm{T}}$, respectively. Given the P echo paths, P adaptive filters are required to cancel the echo signal, and the error signal is given by

$$e[n] = d[n] - \sum_{p=1}^{P}\mathbf{w}_{p,L}^{\mathrm{T}}[n]\mathbf{x}_{p,L}[n] = d[n] - \mathbf{w}_{PL}^{\mathrm{T}}[n]\mathbf{x}_{PL}[n], \tag{26}$$

where $\mathbf{w}_{p,L}[n] = [w_{p,0}[n], \ldots, w_{p,L-1}[n]]^{\mathrm{T}}$ is the p^{th} adaptive filter for the p^{th} echo path and $\mathbf{w}_{PL}[n] = [\mathbf{w}_{1,L}^{\mathrm{T}}[n], \ldots, \mathbf{w}_{P,L}^{\mathrm{T}}[n]]^{\mathrm{T}}$ is the concatenated adaptive filter coefficients vector. Note that here we assume the actual length of the near-end RIRs, K, is different from that of the adaptive filter coefficients vector, L. The minimum mean squared error (MMSE) of $J(\mathbf{w}) = \mathrm{E}\{|\mathrm{e}|^2\}$ for the MCAEC can be obtained by solving the normal equation

$$\mathbf{R}_{\mathbf{x}_{PL}} \mathbf{w}_{PL}[n] = \mathbf{r}_{d\mathbf{x}_{PL}}, \tag{27}$$

where $\mathbf{R}_{\mathbf{x}_{PL}} \equiv \mathrm{E}\{\mathbf{x}_{PL}\mathbf{x}_{PL}^{\mathrm{H}}\}$ and $\mathbf{r}_{d\mathbf{x}_{PL}} \equiv \mathrm{E}\{d\mathbf{x}_{PL}^{*}\}$.

2.2.1 Non-Uniqueness Problem

In MCAEC, the non-uniqueness problem aries due to the highly correlated reference signals that are linearly filtered from the same source. The convergence speed of the LMS algorithm can be slowed down significantly by the non-uniqueness problem. Without loss of generality, Figure 2 shows a diagram of a stereophonic acoustic echo cancellation (SAEC) system where only one of the two microphone signals is analyzed. The non-uniqueness problem [7, 104] can be mathematically characterized as follows.

Let $\mathbf{g}_{p,K}$ be the p^{th} far-end RIR vector of length K, $z_f[n]$ be the far-end source signal, and $z_n[n] = 0$, i.e., no near-end noise. The p^{th} reference signal at sample index n is given by

$$x_p[n] = \mathbf{g}_{p,K}^{\mathrm{T}} \mathbf{z}_{f,K}[n], \quad p = 1, 2. \tag{28}$$

Assuming that the RIRs are linear and time invariant, the two reference signals satisfy the following relationship

$$\mathbf{g}_{2,K}^{\mathrm{T}} \mathbf{x}_{1,K}[n] = \mathbf{g}_{1,K}^{\mathrm{T}} \mathbf{x}_{2,K}[n]. \tag{29}$$

Depending on the length of the adaptive filter, the SAEC can be analyzed as follows.

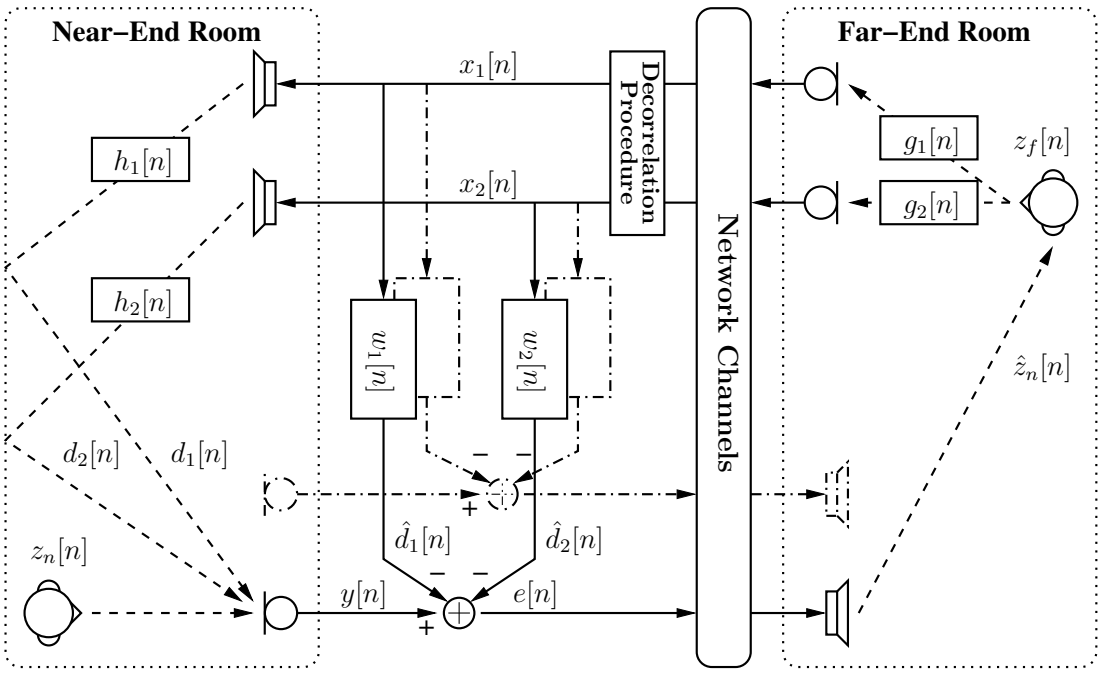


Figure 2: A diagram for stereophonic AEC.

- $L \geq K$: The vector $\mathbf{g}_{\Delta,2L} = [\mathbf{g}_{2,K}^{\mathrm{T}}, \mathbf{0}_{(L-K)\times 1}^{\mathrm{T}}, -\mathbf{g}_{1,K}^{\mathrm{T}}, \mathbf{0}_{(L-K)\times 1}^{\mathrm{T}}]^{\mathrm{T}}$ lies in the null space of $\mathbf{R}_{\mathbf{x}_{2L}}$, since

$$\begin{aligned}\mathbf{R}_{\mathbf{x}_{2L}}\mathbf{g}_{\Delta,2L} &= \begin{bmatrix}\mathbf{x}_{1,L}^{*}[n]\\ \mathbf{x}_{2,L}^{*}[n]\end{bmatrix}\begin{bmatrix}\mathbf{x}_{1,L}^{\mathrm{T}}[n] & \mathbf{x}_{2,L}^{\mathrm{T}}[n]\end{bmatrix}\begin{bmatrix}\mathbf{g}_{2,K}\\ \mathbf{0}_{(L-K)\times 1}\\ -\mathbf{g}_{1,K}\\ \mathbf{0}_{(L-K)\times 1}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{x}_{1,L}^{*}[n]\\ \mathbf{x}_{2,L}^{*}[n]\end{bmatrix}(\mathbf{x}_{1,K}^{\mathrm{T}}[n]\mathbf{g}_{2,K} - \mathbf{x}_{2,K}^{\mathrm{T}}[n]\mathbf{g}_{1,K}) = \mathbf{0}_{2L\times 1}. \qquad (30)\end{aligned}$$

The following set of solutions will satisfy the normal equation (27)

$$\begin{cases}\mathbf{w}_{1,L}[n] = \begin{bmatrix}\mathbf{h}_{1,K}^{\mathrm{T}} + \rho[n]\mathbf{g}_{2,K}^{\mathrm{T}} & \mathbf{0}_{(L-K)\times 1}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\\ \mathbf{w}_{2,L}[n] = \begin{bmatrix}\mathbf{h}_{2,K}^{\mathrm{T}} - \rho[n]\mathbf{g}_{1,K}^{\mathrm{T}} & \mathbf{0}_{(L-K)\times 1}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}\end{cases}, \qquad (31)$$

where ρ is an arbitrary constant.

Therefore, the solution to the MCAEC depends not only on the near-end RIRs but also on the far-end RIRs. The solution has to re-converge when there is

a change in the far-end RIRs, even if the near-end RIRs stay the same. This happens often in a teleconferencing scenario where there are multiple users in one room. Whenever the users in the same room take turns talking, the RIRs in that room change drastically, and thus the convergence speed of the adaptive filter slows down dramatically.

- $L < K$: Since in reality $\mathbf{g}_{1,K}$ and $\mathbf{g}_{2,K}$ are infinite in length, this is normally the case. (29) can be expressed as

$$\mathbf{g}_{2,L}^{\mathrm{T}}\mathbf{x}_{1,L}[n] + \sum_{l=L}^{K-1} g_{2,l}x_1[n-l] = \mathbf{g}_{1,L}^{\mathrm{T}}\mathbf{x}_{2,L}[n] + \sum_{l=L}^{K-1} g_{1,l}x_2[n-l]. \tag{32}$$

  Although $\mathbf{x}_{1,K}$ and $\mathbf{x}_{2,K}$ are linearly related from (29), the same relationship does not hold in general for $\mathbf{x}_{1,L}$ and $\mathbf{x}_{2,L}$ (except for the rare case where $\sum_{l=L}^{K-1} g_{2,l}x_1[n-l] = \sum_{l=L}^{K-1} g_{1,l}x_2[n-l]$). In principle the autocorrelation matrix $\mathbf{R}_{\mathbf{x}_{2L}}$ is full-rank, but in practice it is very ill-conditioned since $\sum_{l=L}^{K-1} g_{2,l}x_1[n-l]$ and $\sum_{l=L}^{K-1} g_{1,l}x_2[n-l]$ are very small. Therefore, for the case where $L < K$, there is a unique solution to the normal equation (27). But the ill-conditioning of the autocorrelation matrix leads to a poor solution due to the high correlation between the reference signals.

2.2.2 Misalignment Problem

The mismatch between the adaptive filter and the actual RIR is quantified by the misalignment, which is defined as

$$\zeta[n] \equiv \frac{\|\mathbf{h}_{PK} - \mathbf{w}_{PK}[n]\|^2}{\|\mathbf{h}_{PK}\|^2}, \tag{33}$$

where near-end RIR is truncated to the same length of the adaptive filter. The relationship between the misalignment and the conditioning of the autocorrelation matrix can be mathematically described as follows.

For the MCAEC, the near-end RIR can be divided into two parts

$$\mathbf{h}_{p,K} = \begin{bmatrix} \mathbf{h}_{p,L} \\ \mathbf{h}_{p,t} \end{bmatrix}, \quad p = 1, 2, \ldots, P, \tag{34}$$

where $\mathbf{h}_{p,t}$ represents the "tail" of the RIR. The microphone signal in the noiseless case can be expressed as

$$y[n] = \sum_{p=1}^{P} \mathbf{h}_{p,L}^{\mathrm{T}} \mathbf{x}_{p,L}[n] + \sum_{p=1}^{P} \mathbf{h}_{p,t}^{\mathrm{T}} \mathbf{x}_{p,t}[n-L] = \mathbf{h}_{PL}^{\mathrm{T}} \mathbf{x}_{PL}[n] + \mathbf{h}_{Pt}^{\mathrm{T}} \mathbf{x}_{Pt}[n-L] \tag{35}$$

where $\mathbf{x}_{p,t}[n-L] = [x_p[n-L], \ldots, x_p[n-N+1]]^{\mathrm{T}}$, $\mathbf{h}_{Pt} = [\mathbf{h}_{1,t}^{\mathrm{T}}, \ldots, \mathbf{h}_{P,t}^{\mathrm{T}}]^{\mathrm{T}}$, and $\mathbf{x}_{Pt} = [\mathbf{x}_{1,t}^{\mathrm{T}}, \mathbf{x}_{2,t}^{\mathrm{T}}, \ldots, \mathbf{x}_{P,t}^{\mathrm{T}}]^{\mathrm{T}}$. The normal equation (27) becomes

$$\begin{aligned} \mathbf{R}_{\mathbf{x}_{PL}} \mathbf{w}_{PL}[n] = \mathbf{r}_{d\mathbf{x}_{PL}} &= \mathrm{E}\{\mathbf{x}_{PL}^{*}[n] d[n]\} \\ &= \mathrm{E}\{\mathbf{x}_{PL}^{*}[n](\mathbf{x}_{PL}^{\mathrm{T}}[n]\mathbf{h}_{PL}) + \mathbf{x}_{PL}^{*}[n](\mathbf{x}_{Pt}^{\mathrm{T}}[n-L]\mathbf{h}_{Pt})\} \\ &= \mathbf{R}_{\mathbf{x}_{PL}} \mathbf{h}_{PL} + \mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}} \mathbf{h}_{Pt}, \end{aligned} \tag{36}$$

where

$$\mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}} \equiv \mathrm{E}\{\mathbf{x}_{PL}[n]\mathbf{x}_{Pt}^{\mathrm{H}}[n-L]\} = \begin{bmatrix} \mathbf{R}_{\mathbf{x}_{1,L}\mathbf{x}_{1,t}} & \mathbf{R}_{\mathbf{x}_{1,L}\mathbf{x}_{2,t}} & \cdots & \mathbf{R}_{\mathbf{x}_{1,L}\mathbf{x}_{P,t}} \\ \mathbf{R}_{\mathbf{x}_{2,L}\mathbf{x}_{1,t}} & \mathbf{R}_{\mathbf{x}_{2,L}\mathbf{x}_{2,t}} & \cdots & \mathbf{R}_{\mathbf{x}_{2,L}\mathbf{x}_{P,t}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{\mathbf{x}_{P,L}\mathbf{x}_{1,t}} & \mathbf{R}_{\mathbf{x}_{P,L}\mathbf{x}_{2,t}} & \cdots & \mathbf{R}_{\mathbf{x}_{P,L}\mathbf{x}_{P,t}} \end{bmatrix}, \tag{37}$$

with $\mathbf{R}_{\mathbf{x}_{i,L}\mathbf{x}_{j,t}} \equiv \mathrm{E}\{\mathbf{x}_{i,L}\mathbf{x}_{j,t}\}$. Assuming that $L < K$ and the autocorrelation matrix $\mathbf{R}_{\mathbf{x}_{PL}}$ is invertible, the solution to (36) becomes

$$\mathbf{w}_{PL}[n] = \mathbf{h}_{PL} + \mathbf{R}_{\mathbf{x}_{PL}}^{-1} \mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}} \mathbf{h}_{Pt}, \tag{38}$$

and the minimized misalignment is given by

$$\zeta_{\min} = \frac{(\mathbf{R}_{\mathbf{x}_{PL}}^{-1} \mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}} \mathbf{h}_{Pt})^{\mathrm{H}} \mathbf{R}_{\mathbf{x}_{PL}}^{-1} \mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}} \mathbf{h}_{Pt}}{\mathbf{h}_{PL}^{\mathrm{H}} \mathbf{h}_{PL}} = \frac{\mathbf{h}_{Pt}^{\mathrm{H}} \mathbf{Q}_{Pt} \mathbf{h}_{Pt}}{\mathbf{h}_{PL}^{\mathrm{H}} \mathbf{h}_{PL}}, \tag{39}$$

where $\mathbf{Q}_{Pt} = \mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}}^{\mathrm{H}} \mathbf{R}_{\mathbf{x}_{PL}}^{-2} \mathbf{R}_{\mathbf{x}_{PL}\mathbf{x}_{Pt}}$. For $L < K$ but large enough L, the following approximations hold

$$\mathbf{R}_{\mathbf{x}_{p,L}\mathbf{x}_{p,t}} \approx \mathbf{0}_{L \times (K-L)}, \quad p = 1, 2, \ldots, P. \tag{40}$$

However, the same relationship does not hold in general for $\mathbf{R}_{\mathbf{x}_{i,L}\mathbf{x}_{j,t}}, i \neq j$, due to the strong correlation between the reference signals. Furthermore, the ill-conditioning of the autocorrelation matrix $\mathbf{R}_{\mathbf{x}_{PL}}$ results in large $\mathbf{Q}_{Pt}$, and hence a large misalignment for $L < K$. Therefore, the misalignment can be still high even if the solution is unique.

As shown in Figure 2, a decorrelation procedure is typically inserted before playing back the reference signal to alleviate the non-uniqueness problem and improve the conditioning of the autocorrelation matrix while introducing minimal distortion to the audio quality and the signal statistics. A handful of inter-channel decorrelation procedures have been proposed in the past to address the non-uniqueness problem. Gansler *et al.* [33] proposed a method that uses a half-wave rectifier to decorrelate the reference signal at the expense of introducing nonlinear distortion. Sugiyama *et al.* [106, 107] proposed an input-sliding technique that introduces one sample delay to disrupt the correlation structure. Other techniques that do not directly distort the signal, such as the phase modulation [55], were also proposed. All of these methods, however, require the introduction of distortion of some sort to the reference signal to solve the non-uniqueness problem in the MCAEC and to maintain a reasonable convergence speed for the LMS based algorithms.

2.3 Robust Acoustic Echo Cancellation

In the previous derivation of the LMS algorithm, the additive noise v was assumed to be absent. Suppose the noise (which could be the near-end voice signal) is added to the microphone signal $y = v + d$, the error signal becomes $e = y - \mathbf{w}^{\mathrm{T}}\mathbf{x}$ and the MSE becomes (omitting the sample index for simplicity)

$$\begin{aligned} J(\mathbf{w}) &= \mathrm{E}\{(y - \mathbf{w}^{\mathrm{T}}\mathbf{x})(y - \mathbf{w}^{\mathrm{T}}\mathbf{x})^*\} \\ &= \sigma_y^2 - \mathbf{r}_{y\mathbf{x}}^{\mathrm{H}}\mathbf{w} - \mathbf{w}^{\mathrm{H}}\mathbf{r}_{y\mathbf{x}} + \mathbf{w}^{\mathrm{H}}\mathbf{R}_{\mathbf{x}}\mathbf{w}. \end{aligned} \tag{41}$$

The gradient of the MSE becomes

$$\begin{aligned}\nabla_{\mathbf{w}^*} J(\mathbf{w}[n]) &= -(\mathbf{r}_{y\mathbf{x}} - \mathbf{R}_{\mathbf{x}}\mathbf{w}[n]) \\ &= -\mathrm{E}\{(v[n] + d[n] - \mathbf{w}^{\mathrm{T}}[n]\mathbf{x}[n])\mathbf{x}^*[n]\} \\ &= -(\mathbf{r}_{v\mathbf{x}} + \mathbf{r}_{b\mathbf{x}}), \end{aligned} \tag{42}$$

with $b = d - \mathbf{w}^{\mathrm{T}}\mathbf{x}$ being the *true*, i.e., noise-free, error. The gradient descent algorithm with noisy update becomes

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu(\mathbf{r}_{v\mathbf{x}} + \mathbf{r}_{b\mathbf{x}}). \tag{43}$$

Note that the optimum solution for the noiseless case was previously obtained by setting $\mathbf{r}_{b\mathbf{x}} = \mathbf{0}_{L\times 1}$ and solving for $\mathbf{w}$. Since the noise is often assumed to be uncorrelated to the reference signal, i.e., $\mathbf{r}_{v\mathbf{x}} = \mathbf{0}_{L\times 1}$, the optimum solution can still be obtained by setting (42) to zero and solving for $\mathbf{w}$. Although this uncorrelatedness assumption is true on average, such a relationship does not always hold when the instantaneous value $\mathbf{r}_{v\mathbf{x}} \approx v\mathbf{x}^*$ is used in the LMS algorithm, where the noisy gradient update becomes

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu(v[n]\mathbf{x}^*[n] + b[n]\mathbf{x}^*[n]), \tag{44}$$

with $v\mathbf{x}^* \neq \mathbf{0}_{L\times 1}$. The noisy gradient may easily cause the adaptive filter to diverge and the optimum solution can no longer be guaranteed in a noisy situation. A conventional solution is to employ a double-talk detector (DTD) such that when the local noise is active (double talk), the adaptation of the filter coefficients is frozen (to avoid the detriment of divergence) and the filter coefficients are updated only during noiseless period.

2.3.1 Double-Talk Detector

One of the earliest DTDs is the Geigel algorithm [25] that is based on the magnitude ratio of the microphone signal to the reference signal. Other methods that are based

on the cross-correlation [6, 35, 62] or the coherence [37, 109, 110] can also be used. However, a false negative in the detection will significantly disrupt the filter adaptation. A more advanced approach, motivated by the robust statistics theory [61], is to apply a compressive nonlinearity to the error signal to limit the sudden fluctuation in the error signal when the DTD fails [4,14,32,36] and to make the adaptive algorithms robust to double talk scenarios.

2.3.2 Error Recovery Nonlinearity

Recently a new generation of RAEC has been proposed based on an integrated system approach without assuming idealized performances of other traditional system components such as DTDs or voice activity detectors (VADs) [114–117, 119]. The algorithm uses error recovery nonlinearity (ERN) and batch adaptation, which allows the adaptive filter to update continuously even during double talk without the use of a DTD. The adaptive filter with ERN is given by

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu\phi(e[n])\mathbf{x}^*[n], \tag{45}$$

where $\phi(\,\cdot\,)$ is a nonlinearity function.

Assuming that the noise v is statistically independent from the true error b, the objective is to derive the optimal nonlinearity that recovers b from the noisy error signal $e = v + b$ through either the MMSE estimate or the maximum *a posteriori* probability (MAP) estimate. A comprehensive list of nonlinearity functions have been derived in [119].

2.3.3 Noise-Robust Adaptive Step-Size

The optimal adaptive step-size for the NLMS algorithm that achieves the largest decrease in the misalignment is [51, 82, 91]

$$\mu_{\text{opt}}[n] = \frac{\mathrm{E}\{|b[n]|^2\}}{\mathrm{E}\{|e[n]|^2\}} = \frac{\mathrm{E}\{|b[n]|^2\}}{\mathrm{E}\{|v[n]|^2\} + \mathrm{E}\{|b[n]|^2\}}, \tag{46}$$

where v and b are assumed to be statistically uncorrelated. Therefore, the optimal step-size rescales the error signal power to be as close to the true error power as possible. Unfortunately, neither the true error b nor the near-end noise v is accessible in real applications.

Alternatively, an effective noise-robust adaptive regularization proposed in [58] and utilized in [114–117, 126, 127] is given by

$$\mathbf{w}[n+1] = \mathbf{w}[n] + \mu e[n] \frac{\|\mathbf{x}[n]\|^2}{\gamma\sigma_v^4 + \|\mathbf{x}[n]\|^4} \mathbf{x}^*[n], \tag{47}$$

where $\gamma > 0$ is a control parameter. A close inspection reveals that the regularization term in (47) is the product of a non-regularized normalization factor and a Wiener-like scaling factor [117]

$$\frac{\|\mathbf{x}[n]\|^2}{\gamma\sigma_v^4 + \|\mathbf{x}[n]\|^4} \approx \frac{1}{\|\mathbf{x}[n]\|^2} \frac{\sigma_b^2}{\gamma\Big(\frac{\mathrm{E}\{\|\mathbf{h}-\mathbf{w}[n]\|^2\}}{L^2}\frac{\sigma_v^2}{\sigma_x^2}\Big)\sigma_v^2 + \sigma_b^2}. \tag{48}$$

The near-equality is obtained by approximating $\|\mathbf{x}\|^2 = L\sigma_x^2$ for a long filter length $L \gg 1$ and assuming $\mathbf{x}$ is white and statistically independent from $\mathbf{h}$ such that $\sigma_b^2 = \mathrm{E}\{\|\mathbf{h} - \mathbf{w}\|^2\}\sigma_x^2$. (48) ensures that the Wiener step-size control is carried out properly when the adaptive filter has not reached sufficient convergence (i.e., for large $E\|\mathbf{h} - \mathbf{w}\|^2$) or when the mixing system is weakly excited (i.e., for small σ_x^2/σ_v^2). In this sense, (48) also performs the VAD on the reference signal and adaptively controls the step-size to ensure a stable adaptation during the weakly-excited situation. Since accurate estimation and tracking of σ_v^2 are very difficult, $\sigma_v^2 \approx \sigma_e^2$ is used instead in (47) for the actual implementation.

2.4 Residual Echo and Noise Suppression

Due to the modeling mismatch (misalignment) between the adaptive filter and the near-end RIR, there will inevitably be residual echo that can not be fully canceled by the AEC alone. The residual echo suppression (RES) based on single-channel noise reduction techniques is often employed after the AEC to further suppress the

echo. The single-channel noise reduction problem can be mathematically formulated as follows.

From the perspective of the RES, the output of the AEC $e = v + b$ contains the desired near-end speech signal v plus the uncorrelated residual echo b. The noisy observed signal e is transformed into the time-frequency domain by the short-time Fourier transform (STFT) given by

$$E_k[m] = \sum_{n=0}^{N-1} e[n + mR] w_{\mathrm{A}}[n] \omega_N^{kn}, \tag{49}$$

where k is the frequency index, m is the frame index, N is the frame size, R is the frame shift size, w_{A} is an analysis window of size N (e.g., Hanning window), and $\omega_N = \exp(-j\frac{2\pi}{N})$. Given an estimate of the clean speech STFT $\hat{V}[k, m]$, an estimate of the clean speech signal is obtained by applying the inverse shoft-time Fourier transform (ISTFT)

$$\hat{v}[n] = \sum_{m} \sum_{k=0}^{N-1} \hat{V}_k[m] w_{\mathrm{S}}[n - mR] \omega_N^{-k(n-mR)}, \tag{50}$$

where w_{S} is a synthesis window that is biorthogonal to the analysis window w_{A}. For perfect reconstruction of a signal, the analysis and synthesis windows must satisfy the so-called *completeness condition*, i.e.,

$$\sum_{m} w_A[n + mR] w_S[n + mR] = 1, \quad \forall n. \tag{51}$$

2.4.1 Wiener Filter

The objective of the single-channel noise reduction is to find an estimator $\hat{V}_k[m]$ that minimizes the conditional expectation of a distortion measure, given a set of noisy measurements

$$\hat{V}_k[m] = \arg\min_{\hat{V}} \mathrm{E}\{\mathcal{D}(V_k[m], \hat{V}) | E_0[m], E_1[m], \dots\}, \tag{52}$$

where $\mathcal{D}(V_k, \hat{V}_k)$ is the distortion measure between V_k and $\hat{V}_k$. This is usually done by multiplying the noisy spectral component E_k by a non-negative and real-valued

suppression gain G_k to obtain the estimate $\hat{V}_k$ of the desired signal

$$\hat{V}_k[m] = G_k[m]E_k[m], \tag{53}$$

where the noisy spectral components in each frequency bin are assumed to be statistically independent and the estimator can be derived from E_k only. The frame index m from now on will be omitted for simplicity unless otherwise mentioned.

The squared-error distortion, defined as

$$\mathcal{D}_{\mathrm{SE}}(V_k, \hat{V}_k) \equiv |V_k - \hat{V}_k|^2, \tag{54}$$

is a commonly used criterion which leads to the optimization of the following MSE

$$\begin{aligned} \mathrm{E}\{\mathcal{D}_{\mathrm{MSE}}(V_k, \hat{V}_k)\} &= \mathrm{E}\{|V_k - G_k E_k|^2\} = \mathrm{E}\{|V_k - G_k(V_k + B_k)|^2\} \\ &= (1 - G_k)^2 \lambda_V[k] + G_k^2 \lambda_B[k], \end{aligned} \tag{55}$$

where the desired signal and the residual echo are modeled as statistically uncorrelated complex Gaussian random variables with zero mean and the cross-terms $\mathrm{E}\{V_k B_k^*\} = \mathrm{E}\{V_k^* B_k\} = 0$. The desired signal and the residual echo variances in the k^{th} frequency bin are given by $\lambda_V[k] \equiv \mathrm{E}\{|V_k|^2\}$ and $\lambda_B[k] \equiv \mathrm{E}\{|B_k|^2\}$, respectively. (55) can be minimized by taking the partial derivative with respective to G_k and equating the partial derivative to zero, i.e.,

$$\frac{\partial \mathrm{E}\{\mathcal{D}_{\mathrm{MSE}}(V_k, \hat{V}_k)\}}{\partial G_k} = -2(1 - G_k)\lambda_V[k] + 2G_k\lambda_B[k] = 0. \tag{56}$$

The MMSE in (55) is obtained by solving for G_k in (56) and the solution is given by

$$G_k^{\mathrm{W}} = \frac{\lambda_V[k]}{\lambda_V[k] + \lambda_B[k]} = \frac{\xi_k}{\xi_k + 1}, \tag{57}$$

which is the most famous frequency-domain *Wiener filter*. The term $\xi_k \equiv \lambda_V[k]/\lambda_B[k]$ is defined as the *a priori* signal-to-noise ratio (SNR).

2.4.2 Log-Spectral Amplitude Estimator

Another useful distortion measure is the log-spectral amplitude (LSA) distortion that is given by

$$\mathcal{D}_{\text{LSA}}(V_k, \hat{V}_k) = (\log|V_k| - \log|\hat{V}_k|)^2, \tag{58}$$

and minimizing the MSE with respect to $\mathcal{D}_{\text{LSA}}$ leads to the suppression gain [30]

$$G_k^{\text{LSA}} = G_k^{\text{W}} \exp\left(\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right), \tag{59}$$

where $\nu_k \equiv G_k^{\text{W}}\gamma_k$ and γ_k is the *a posteriori* SNR defined as $\gamma_k \equiv |E_k|^2/\lambda_B[k]$. Perceptually speaking, (59) has much less musical noise than the Weiner filter [53], thus making it particularly popular throughout speech enhancement literature.

2.4.3 A Priori Signal-to-Noise Ratio Estimator

Given the suppression gain (57) and (59), additional estimators are still required to suppress the noise, i.e., the *a priori* SNR estimator and the noise power estimator. The *a priori* SNR estimator can be derived using the maximum likelihood (ML) estimation [29], which is based on the estimation of signal variance by maximizing the joint conditional probability density function (PDF) of M consecutive observation $\mathcal{E}_k[m] \equiv \{E_k[m], E_k[m-1], \ldots, E_k[m-M+1]\}$ given λ_V and λ_B, i.e.,

$$\hat{\lambda}_{V,\text{ML}}[k] = \arg\max_{\lambda_V} p(\mathcal{E}_k[m]|\lambda_V[k], \lambda_B[k]). \tag{60}$$

Based on the Gaussian statistical model and the statistical independence assumption for the spectral components across different frames, the conditional PDF is given by

$$p(\mathcal{E}_k[m]|\lambda_V[k], \lambda_B[k]) = \prod_{l=0}^{M-1} \frac{1}{\pi(\lambda_V[k] + \lambda_B[k])} \exp\left(-\frac{|E_k[m-l]|^2}{\lambda_V[k] + \lambda_B[k]}\right). \tag{61}$$

Therefore, the ML estimator $\hat{\lambda}_V$ of λ_V is equal to

$$\hat{\lambda}_V[k] = \begin{cases} \frac{1}{M}\sum_{l=0}^{M-1}|E_k[m-l]|^2 - \lambda_B[k], & \text{if non-negative,} \\ 0, & \text{otherwise.} \end{cases} \tag{62}$$

Note that the estimator $\hat{\lambda}_V$ is constrained to be non-negative. The ML estimator for the *a priori* SNR is given by

$$\hat{\xi}_{k,\mathrm{ML}}[m] = \begin{cases} \frac{1}{M}\sum_{l=0}^{M-1}\gamma_k[m-l]-1, & \text{if non-negative,} \\ 0, & \text{otherwise.} \end{cases} \tag{63}$$

The actual implementation is a recursive average given by

$$\bar{\gamma}_k[m] = \alpha_{\mathrm{ML}}\bar{\gamma}_k[m-1] + (1-\alpha_{\mathrm{ML}})\frac{\gamma_k[m]}{\beta_{\mathrm{ML}}}, \tag{64}$$

$$\hat{\xi}_{k,\mathrm{ML}}[m] = \max\{\bar{\gamma}_k[m]-1, 0\}. \tag{65}$$

where $0 \ll \alpha_{\mathrm{ML}} < 1$ is a forgetting factor and $\beta_{\mathrm{ML}} \geq 1$ is a correction factor.

One drawback of the ML *a priori* SNR estimator is that the residual noise after suppression becomes the annoying "musical noise," which arises due to the over-suppression of the spectral components in some frequency bins, resulting in metallic and unnatural sound after suppression. Another *a priori* SNR estimator is the decision-directed (DD) approach [29], which is given by

$$\hat{\xi}_{k,\mathrm{DD}}[m] = \alpha_{\mathrm{DD}}\frac{|\hat{V}_k[m-1]|^2}{\lambda_B[k]} + (1-\alpha_{\mathrm{DD}})\max\{\gamma_k[m]-1, 0\}, \tag{66}$$

where $0 \ll \alpha_{\mathrm{DD}} < 1$ is a forgetting factor. The name "decision-directed" comes from the fact that the *a priori* SNR estimator is updated based on the amplitude estimation of the previous frame. Compared to the ML *a priori* SNR estimator, the DD *a priori* SNR estimator is generally more preferable since the residual noise using DD is relatively smoother and more pleasant [29].

2.4.4 Noise Power Estimator

For a stationary type of noise, e.g., car noise or air conditioner noise, several noise power estimation algorithms, such as the minimum statistics [85,86], the minima controlled recursive averaging method [19, 20], or the MMSE-based noise power estimation [39–41,54], can be used. However, the residual echo resulted from the AEC is usually highly colored and nonstationary. Other methods such as the equivalent transfer

function based method [47,49] or the coherence function based method [26–28,48,113] can be used for residual echo power estimation.

The regressed-based method [9,17] models the magnitude of the short-term spectrum of the residual echo in terms of the magnitudes of the current and previous frames of the reference signal. This is based on the assumption that the AEC is able to model and cancel the effect of the relatively early echoes. The residual echo can reasonably be assumed to contain a part of the early echo and most of the late reverberation. Since the AEC captures a significant part of the phase information, only the magnitude of the reference signal is used to model the residual echo. The advantage of this model is that the regression coefficients can be easily estimated and tracked using any adaptive algorithm.

2.5 Performance Measures

We have previously defined the misalignment for the AEC as

$$\text{Misalignment (dB)} \equiv 10\,1\log_{10}\left(\frac{\|\mathbf{h}-\mathbf{w}[n]\|^2}{\|\mathbf{h}\|^2}\right), \tag{67}$$

which shows the system identification performance. Another useful measure for the AEC is the echo return loss enhancement (ERLE)

$$\text{ERLE (dB)} \equiv 10\log_{10}\left(\frac{\sum_n |y[n]|^2}{\sum_n |e[n]|^2}\right), \tag{68}$$

which measures the degree of "enhancement" of the microphone signal in terms of how much the echo signal is canceled. For a well performing AEC, the ERLE should be as high as possible, while the misalignment should be as low as possible.

However, when the additive noise v is not zero, it is more useful to use the true echo return loss enhancement (TERLE)

$$\text{TERLE (dB)} \equiv 10\log_{10}\left(\frac{\sum_n |y[n]-v[n]|^2}{\sum_n |e[n]-v[n]|^2}\right) = 10\log_{10}\left(\frac{\sum_n |d[n]|^2}{\sum_n |b[n]|^2}\right), \tag{69}$$

i.e., the ERLE measured after the near-end noise is subtracted from both the microphone signal and the error signal. The reason for using TERLE is that most often v

contains the near-end speech signal that is to be consumed by the far-end listeners. An over-suppression of the residual echo by RES, i.e., suppressing the residual echo as well as the near-end speech, may result in high ERLE but actually low TERLE. On one extreme, the error signal after the AEC is heavily suppressed, i.e., $e \approx 0$, the ERLE will be very high, but the TERLE will stay low since the near-end signal is taken into account in the denominator of TERLE.

For measuring the RES performance, the following three measures are used: the segmental signal-to-noise ratio (SSNR), the log-spectral distortion (LSD), and the Performance Evaluation of Speech Quality (PESQ). The SSNR is defined as

$$\text{SSNR (dB)} \equiv \frac{1}{\mathcal{J}} \sum_{m=0}^{\mathcal{J}-1} \mathcal{T}\left\{ 10 \log_{10} \frac{\sum_{n=0}^{N-1} v^2[n+mR]}{\sum_{n=0}^{N-1} (v[n+mR] - \hat{v}[n+mR])^2} \right\}, \tag{70}$$

where $\mathcal{J}$ is the number of frames, N is the frame size, $\mathcal{T}\{\,\cdot\,\}$ confines the SNR at each frame to perceptually meaningful range between -10 dB and 35 dB. The LSD is defined as

$$\text{LSD (dB)} \equiv \frac{1}{\mathcal{J}} \sum_{m=0}^{\mathcal{J}-1} \left\{ \frac{1}{N/2+1} \sum_{k=0}^{N/2} \left[10 \log_{10} \left(\frac{\mathcal{C}\{V_k[m]\}}{\mathcal{C}\{\hat{V}_k[m]\}} \right) \right]^2 \right\}^{\frac{1}{2}}, \tag{71}$$

where $\mathcal{C}\{V_k[m]\} \equiv \max\{|V_k[m]|^2, \delta\}$ is the clipped spectral power such that the log-spectrum dynamic range is confined to 50 dB, i.e., $\delta \equiv 10^{\frac{-50}{10}} \max_{k,m}\{|V_k[m]|^2\}$. Note that the SSNR measures the accuracy of the estimated desired signal in the time domain, while the LSD measures it in the frequency domain.

The PESQ [71, 97] is an objective measurement tool that predicts the results of the mean opinion score (MOS) in subjective listening tests and is a useful measure for assessing the speech quality that cannot be fully addressed by the SSNR or the LSD alone. For example, a simple delay of a signal drastically lowers the SSNR while the perceptual quality before and after the delay is essentially the same.

CHAPTER III

SYSTEM APPROACH TO ACOUSTIC ECHO CANCELLATION AND RESIDUAL ECHO SUPPRESSION

In this chapter, we present several methods to systematically combine the acoustic echo cancellation (AEC) and residual echo suppression (RES) to improve the echo reduction results. Traditionally, the AEC and the RES are developed with different criteria and are treated as two separate units. The two units normally operates in their own framework that has no commonality or shared information between them. Through the system approach, the two units are combined in such a way that the performance of one of the unit can be further enhanced given the processed signal from the other unit. In Section 3.1, we discuss the system approach to RES, where the RES unit can be enhanced by using the knowledge of the linearly estimated echo signal from the AEC. In Section 3.2, we take the other direction and utilize the enhanced speech signal from the RES unit to boost the performance of the AEC. Finally, we show how the system approach works in the *Kinect*$^{\text{TM}}$ audio system to demonstrate the benefit of the system approach.

3.1 System Approach to Residual Echo Suppression

The robust acoustic echo cancellation (RAEC) described in Section 2.3 and proposed in [117] allows the adaptive filter to update continuously even during double talk without the use of a double-talk detector (DTD) or a voice activity detector (VAD). This *robust* AEC setup warrants a new perspective for the problem of RES. Due to natural mismatches between the room impulse response (RIR) and the adaptive filter,

the actual echo cannot be cancelled perfectly by the AEC echo estimate. To ensure high-quality telephony, the remaining echo must be further suppressed by RES, which requires a residual echo estimate. However, the residual echo is often corrupted by noise, e.g., near-end speech, and can be difficult to estimate accurately.

A new residual echo estimation method that exploits the *nonlinearly estimated echo* by the log-spectral amplitude (LSA) estimator [30] and the *linearly estimated echo* by an adaptive filter of the AEC was proposed in [126]. The procedure results in a very close representation of the noise-free residual echo. Echo cancellation and echo suppression are two distinct processes, where echo cancellation *subtracts* the estimated echo samples from the microphone signal. It usually introduces much less distortion compared to echo suppression, which *attenuates* the signal amplitude. Traditional RES techniques based on frequency-domain Wiener filtering are sensitive to the accuracy of estimated signal-to-noise ratio (SNR) and may introduce near-end speech distortion or musical noise [81]. RES ideally should produce minimal distortion of the near-end signal during both single talk and double talk. Towards this end a psychoacoustic postfilter [48] is used in this section to suppress the residual echo as much as possible without introducing audible distortion to the near-end speech. The overall goal is to achieve a system combination of individually designed components that together facilitate an improved performance of the AEC system as a whole.

3.1.1 Psychoacoustic Postfilter

Based on the additive noise model $e = v + b$, an LSA estimator is often used for RES to estimate v. However, using the LSA estimator with the decision-directed (DD) *a priori* SNR estimator requires a residual echo variance estimate $\hat{\lambda}_B[k]$, which is one of the most critical parts that influence the RES performance. Given a residual echo variance estimate, we can express the LSA filter as a nonlinear function

$$\tilde{V} = f_{\mathrm{LSA}}\{E, \hat{\lambda}_B\}. \tag{72}$$

Due to the suppressive nature of the LSA gain, i.e., $0 \leq G_k^{\mathrm{LSA}} \leq 1$, the near-end signal can potentially be distorted when the residual echo magnitude is attenuated too much. However, during periods of high background noise levels or double talk, less suppression is required since the residual echo will be masked by the near-end signal [48, 49]. By incorporating this frequency masking property, the psychoacoustic postfilter is derived as follows. Generally, the near-end signal is estimated in the frequency domain as

$$\hat{V}_k = H_k E_k = H_k(V_k + B_k). \tag{73}$$

Assuming that the near-end signal and the residual echo are statistically uncorrelated, the overall distortion of the near-end signal can be written as

$$\mathrm{E}\{|V_k - \hat{V}_k|^2\} = (1 - H_k)^2\,\mathrm{E}\{|V_k|^2\} + H_k^2\,\mathrm{E}\{|B_k|^2\}, \tag{74}$$

where the second term represents the distortion of the residual echo. To minimally impact the near-end speech, a minimum level of suppression is chosen such that the residual echo distortion equals the masking threshold $T_V[k]$ of the near-end signal. The psychoacoustic postfilter gain is given by [48]

$$H_k = \min\left\{1, \sqrt{\frac{T_V[k]}{\lambda_B[k]}}\right\}. \tag{75}$$

Therefore, if the residual echo is already masked by the near-end signal, i.e., $T_V[k] > \lambda_B[k]$, the psychoacoustic postfilter gain will be set to 1, and the near-end signal will be undistorted.

A block diagram of the psychoacoustic postfilter is shown in Figure 3. The operation of the postfilter is as follows [48]:

- Obtain a residual echo estimate $\hat{B}_k$.
- Apply the LSA gain (59) to E_k by using $\hat{B}_k$ to obtain a rough estimate of the near-end signal $\tilde{V}_k$.

- Calculate the masking threshold $T_V[k]$ by using $\tilde{V}_k$.
- Calculate the postfilter gain based on (75) and apply it to E_k to obtain a better near-end signal estimate $\hat{V}_k$.

Since both the near-end signal and the residual echo are unknown, the problem of obtaining an accurate residual echo estimate remains.

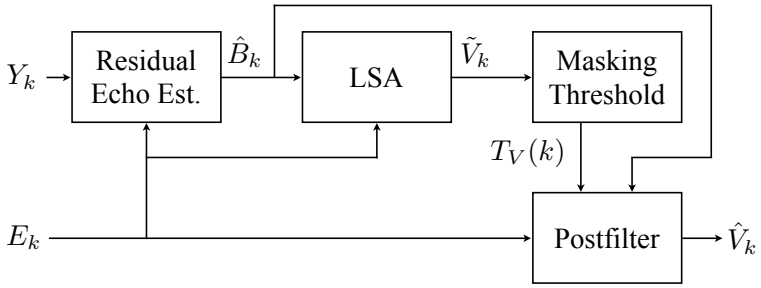


Figure 3: A block diagram of the psychoacoustic postfilter.

3.1.2 Residual Echo Estimation Method

The near-end microphone signal is modeled as

$$y[n] = d[n] + v[n], \tag{76}$$

which contains the true echo d and the near-end signal v. We first estimate d by treating v as an additive noise to be removed from y by LSA filtering, i.e., $\tilde{D} = f_{\text{LSA}}\{Y, \lambda_V\}$. The instantaneous estimate of $\lambda_V[k]$ is obtained from the output of the RAEC, i.e., $e = v + b$, as we assume $e \approx v$ after the convergence of the adaptive filter, or at least $|E_k| \approx |V_k|$ due to the sparsity of a speech signal in the frequency domain. By applying the LSA filter to Y_k, D_k will be emphasized whereas V_k will be suppressed. Finally, the difference between the nonlinear echo estimate provided by the LSA filter and the linear echo estimate provided by the AEC closely represents the true residual echo:

$$\hat{b}[n] = \tilde{d}[n] - \hat{d}[n]. \tag{77}$$

That is, nonlinear processing by the LSA filter should not alter the residual echo contained in $d = b + \hat{d}$, since b simply represents any remaining part of d that cannot be cancelled linearly by adaptive filtering. Other interpretations are as follows.

The basic assumption is that due to the noise-robustness of a combination of AEC and error recovery nonlinearity (ERN), the signal power λ_B is small compared to λ_D during single talk or λ_V during double talk after the adaptive filter has converged. Assuming that v contains speech only and is free from the background noise, analysis of the LSA filtering can be categorized into the three cases below:

- Near-end talk (NT): $D_k = 0$, and only the near-end signal is active, i.e., $Y_k = E_k = V_k$. The LSA filter will suppress all near-end signals and $\tilde{D}_k \approx 0$.

- Single talk (ST): $V_k = 0$, and only the far-end talker is active, i.e., $Y_k = D_k$ and $E_k = B_k$. Since $\lambda_D[k] \gg \lambda_B[k]$, the LSA estimator will operate in high SNR mode. Therefore, $G_{\text{LSA}} \approx 1$, and the LSA filter will not attenuate Y_k and output $\tilde{D}_k \approx D_k$.

- Double talk (DT): Both near-end talker and far-end talker are active, i.e., $Y_k = D_k + V_k$ and $E_k = V_k + B_k$. Since $\lambda_V[k] \gg \lambda_B[k]$ (as a result of the effective RAEC), and based on the assumption that V_k and B_k are zero mean and statistically uncorrelated random variables, we can write

$$\begin{aligned} \lambda_E[k] &= \mathcal{E}\{|E_k|^2\} = \mathrm{E}\{|V_k + B_k|^2\} \\ &= \mathrm{E}\{|V_k|^2\} + \mathrm{E}\{|B_k|^2\} \\ &= \lambda_V[k] + \lambda_B[k] \approx \lambda_V[k]. \end{aligned} \tag{78}$$

  Therefore, the LSA filter will reduce mostly the near-end signal contained in Y_k, hence $\tilde{D}_k \approx D_k$.

Spectrograms of the reference signal X, the AEC output E, the true residual echo B, and the proposed residual echo estimate $\hat{B}$ are shown in Figure 4. For

clarity, the spectrum of only up to 4 kHz is shown since a speech signal is mostly concentrated around low frequencies. A 10 dB segmental signal-to-noise ratio (SSNR) air conditioner noise is added to the microphone signal. The figure shows that E contains the near-end speech, the air conditioner noise, and the residual echo. We note that due to the strong disturbance from V during double talk, $\tilde{D}$ may not be accurate enough and $\hat{B}$ is possibly overestimated. However, the masking threshold will also be high during double talk, and overestimation of $\hat{B}$ will not pose a problem in such a case. On the other hand, the near-end signal contains the air conditioner noise during single talk, and V is not strictly equal to zero. Then $\hat{B}$ will contain the true residual echo as well as some background noise. Nevertheless, we will show in Section 3.3 that this minor disturbance to our residual echo estimate only slightly affect the overall system performance.

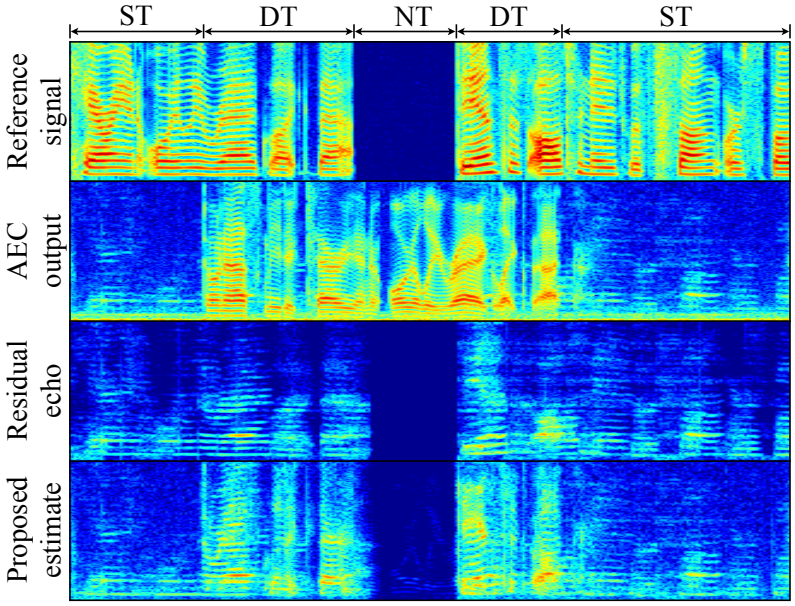


Figure 4: Spectrograms comparing the proposed residual echo estimate to the true residual echo.

3.2 System Approach to Acoustic Echo Cancellation

The RAEC system uses an ERN and batch-wise adaptation to permit the adaptive filter to update continuously even during double talk with robustness to mis-estimation of the signal statistics. However, the signal models for ERN are fixed in previous approaches. Since the underlying assumptions may change as the near-end and far-end signal distributions change, the estimated statistics may not truly reflect the best possible system performance. Motivated by the positive results from the system approach to RES, we continue the investigation to further enhance the entire system performance.

An improved system approach to RAEC that utilizes the postfiltered output to further assist the RAEC system was presented in [127]. Specifically, the RAEC system, shown as a component within the combined AEC system in Figure 5, uses the output from postfiltering to form an integrated loop that boosts the overall AEC performance. We also keep track of the RAEC output that is produced both with and without postfiltering, and the final output is chosen from the best output among the two.

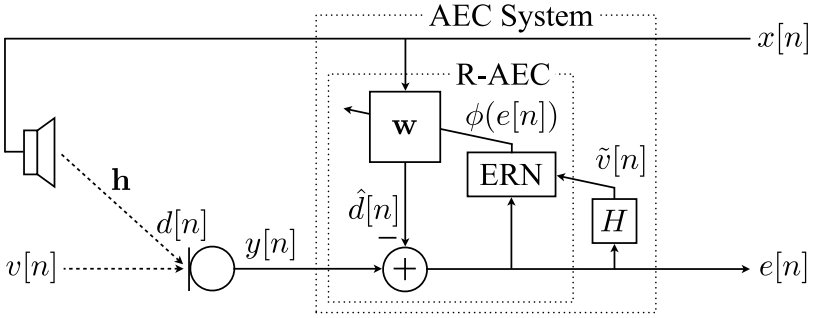


Figure 5: The system approach to AEC with an adaptive filter $\mathbf{w}$, an error recovery nonlinearity, and a postfilter H that directly assists the RAEC component (a separate postfilter for RES is omitted).

Traditionally, a postfilter is used for RES to further suppress the residual echo before sending the near-end signal estimate out to the far end. In our proposed system, however, another postfiltering process is used to directly assist the ERN

component such that the signal statistics of the near-end speech and the residual echo are more accurately estimated through block-iterative adaptation of the whole system. The resulting estimation error from the RAEC system will thus contain less residual echo by incorporating information from the postfilter.

The postfilter used here is a simplified version of the postfilter presented in Section 3.1.1, since only a rough estimate of the near-end signal $\tilde{v}$ is required. The postfilter is expressed in terms of the LSA filter as

$$\tilde{V} = f_{\text{LSA}}\{E, \hat{\lambda}_B\}. \tag{79}$$

For the system approach to RES, the residual echo estimate required for LSA is obtained by first estimating the true echo d from the microphone signal y and the RAEC output e via LSA as

$$\tilde{D} = f_{\text{LSA}}\{Y, \lambda_V\} \approx f_{\text{LSA}}\{Y, \lambda_E\}, \tag{80}$$

and subsequently by

$$\hat{b}[n] = \tilde{d}[n] - \hat{d}[n]. \tag{81}$$

3.2.1 System Approach to Error Recovery Nonlinearity

Several choices of nonlinearity have been discussed and compared in [117, 119]. The nonlinearity that produces the best results is used in this work. Assuming that the residual echo b is Gaussian and the near-end signal v is Laplace distributed, we can derive the optimal nonlinearity function based on the maximum *a posteriori* probability (MAP) estimate [119]

$$\phi_{\text{MAP}}^{\text{GL}}(e) = \begin{cases} \text{sign}(e)t, & |e| \geq t = \sigma_b^2/\alpha_v, \\ e, & \text{otherwise}, \end{cases} \tag{82}$$

where σ_b^2 is the variance of b and α_v is the scale parameter of v. (82) is much simpler to implement than the minimum mean squared error (MMSE) nonlinearity used in [117] while providing practically the same AEC performance [119].

Let the SNR of the residual echo and the near-end speech be defined as $\xi = \sigma_b^2/\alpha_v^2$ and rewrite the threshold in (82) as $t = \xi\alpha_v$. Then by assuming that the adaptive filter has converged sufficiently (i.e., $e \approx v$) and letting $\xi \approx 1$, the threshold can be approximated in the frequency domain as

$$T_1[k] = \sqrt{S_{ee}[k]}/\eta_1, \tag{83}$$

where η_1 is an over-suppression factor and S_{ee} is the power spectrum of the RAEC output e. S_{ee} is given by

$$S_{ee,i}[k] = \beta S_{ee,i-1}[k] + (1-\beta)|E_i[k]|^2, \tag{84}$$

where i is the iteration index and β is a smoothing factor.

By (83), which is the simplified threshold estimate actually used in [117, 119] for $\eta_1 = 1$, (82) roughly mimics the double-talk robust nonlinearity derived through the robust statistics theory [119]. The threshold in (82) can also be rewritten as $t = \sqrt{\xi}\sigma_b \approx \sigma_e$ for $\xi \approx 1$. Then, although S_{ee} indeed approximates S_{bb} well during single talk with minimal background noise, it overestimates S_{bb} during double talk due to the near-end speech, which may consequently lead to adaptation instability. This is why the overall AEC performance, especially during double talk, is shown in [118] to improve when the over-suppression factor and double-talk detection (to reduce the step-size by a factor of 1/2 during double talk) are included.

We can obtain a more accurate estimate of the threshold by

$$T_2[k] = \frac{S_{bb}[k]}{\eta_2\sqrt{S_{ee}[k]/2}}, \tag{85}$$

along with η_2 for over-suppression and S_{ee} to estimate S_{vv} after sufficient convergence. The $1/\sqrt{2}$ factor compensates for the overestimation of the scale parameter by $\sqrt{S_{ee}}$ since the variance of v is equal to $2\alpha_v^2$. The power spectrum of the residual echo S_{bb}

is estimated by

$$S_{bb,i}[k] = \beta S_{bb,i-1}[k] + (1-\beta)|\tilde{B}_i[k]|^2, \tag{86}$$

$$\tilde{B}_i[k] = E_i[k] - \tilde{V}[k]. \tag{87}$$

As discussed above, the original threshold estimate in (83) was formed in relation to the compressive nonlinearity based on the robust statistics, where σ_b is represented well by σ_e during single talk. However, it lacks the adaptation of the scaling term that further enhances the robustness to impulsive noise [119], hence the estimate of σ_e can easily overestimate σ_b during double talk due to the "burstiness" of a speech signal. On the other hand, the proposed threshold estimate of (85) has S_{ee} in the denominator, which safeguards against the effect of overestimation by scaling down the step-size in such a case to ensure adaptation stability. Furthermore, it has explicit dependence on S_{bb} and S_{ee} such that there is much tighter coupling between the error enhancement process and the adaptive filtering than previously. That is, better estimation of S_{bb} for the ERN will subsequently lead to better estimation of S_{ee} by the adaptive filter.

We note that (81), as was presented in Section 3.1.2, does not work as well when used to directly estimate S_{bb} in (85), although it is still used indirectly to obtain $\tilde{V}[k]$ in (86). We have experimentally verified that (85) is better controlled via (86) rather than (81). This may be because the threshold estimate given by (85) is linearly dependent on the direct output RAEC, i.e., e, which contains information from both b and v. The linear relationship is maintained better by (86) to provide more beneficial interaction between the adaptation process and the threshold control than freely using nonlinearly estimated $\hat{b}$ of (81). The explanation is in line with the prior observations in [117, 119] that an adaptive algorithm should be able to converge naturally to the best solution in the presence of near-end distortion if the stability is consistently maintained, and that smoother tracking of the signal statistics is more effective than

otherwise. Even though the rough near-end speech estimate from LSA may introduce more signal distortion when compared to the perceptually-masked postfiltering used in Section 3.1.1, it is sufficient for assisting the ERN component through smoothing by (86).

3.2.2 Two-Pass Adaptation

The near-end speech estimate $\tilde{v}$ obtained from LSA is possible when (80) indeed approximates the true echo. This is true when the adaptive filter has converged, i.e., when $\lambda_E \approx \lambda_V$ as in Section 3.1.2, which nonetheless cannot be guaranteed at all time. Therefore, the overall system is divided into a two-pass system. In the first pass, the threshold of the ERN in (82) is first estimated using (83). Once we obtain the rough near-end signal estimate, S_{bb} is calculated from (86) to update the threshold via (85). After the first pass, the adaptive filter coefficients are saved. In the second pass, the RAEC system is re-adapted from a whole new set of parameters. In particular, the filter coefficients in the second pass are adapted using the proposed threshold (85) independent of those in the first pass. We can view the first pass and the second pass as two separate RAEC units with different threshold values to obtain the best estimation error possible. The two-pass system is necessary since the effectiveness of the proposed threshold is contingent upon the residual echo estimate, which is better obtained when the adaptive filter is *sufficiently converged*.

3.2.3 Hybrid Approach

The proposed threshold may perform differently at certain frequency locations, e.g., it works better mainly at low frequencies. Therefore, a frequency selection is applied to the RAEC outputs from the first and the second passes as follows. Let us denote the first-pass RAEC output as e_1 and the second-pass one as e_2. Then the overall

system output is given by

$$E[k] = \begin{cases} E_2[k], & \text{double talk,} \\ |E_{\min}[k]|e^{j\psi\{E_2[k]\}}, & \text{otherwise,} \end{cases} \tag{88}$$

where $\psi\{E_2[k]\}$ is the phase of $E_2[k]$ (experimental results show that the phase of either E_1 or E_2 may be used) and

$$|E_{\min}[k]| = \min\{|E_1[k]|, |E_2[k]|\}, \tag{89}$$

which is determined during single talk. The advantage of the frequency selection is that the error consists mostly of the residual echo during single talk. Thus by selecting the minimum output amplitude across all frequencies, the effect of the remaining noise is further minimized. On the other hand, the residual echo is dominated largely by the near-end speech during double talk. A frequency selection of the error signal in such a case results in relatively the same amount of the residual echo at the output. Therefore, only the output with potentially the least amount of the residual echo, i.e., the error signal from the second pass RAEC, is chosen.

The overall system based on the proposed ERN threshold, the two-pass adaptation, and the hybrid approach is shown in Figure 6. Note that a DTD is used for the output selection only and not for the entire RAEC block, as the RAEC system can adapt continuously during double-talk situations. The application of DTD allows our system to identify different signal mixing environments and facilitates the output selection process to obtain a result with the lowest amount of residual echo, which is quite different from the traditional use of DTD. Thus a DTD can still be utilized effectively in such a manner through the system approach.

3.3 Experimental Evaluation

3.3.1 System Approach to Residual Echo Suppression

16 kHz 16-bit PCM recordings of female and male speech signals from the TIMIT database were used as the far-end and the near-end signals, respectively. The far-end

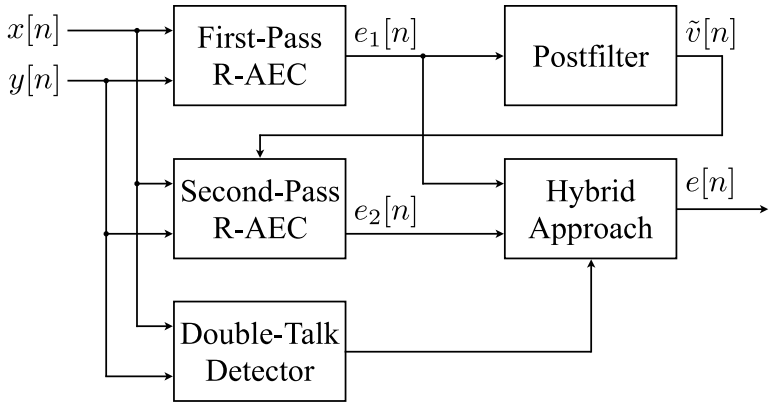


Figure 6: A block diagram of the overall AEC system based on the proposed ERN threshold, the two-pass adaptation, and the hybrid approach.

signal was normalized to $[-1, 1]$ range, and the echo signal was re-scaled to produce a 10 dB echo return loss before the addition of the male speech of equal power. To simulate the real world situation, the far-end and the near-end speakers took turns talking with an overlap of 1 second. Air conditioner noise at SSNR of 0 to 30 dB with 10 dB increments were added to the microphone signal. 10 test pairs of near-end signal and far-end signal with an average length of 20 seconds were created. The first 5 seconds of each test pair contained no near-end speech to insure convergence. These segments were removed prior to evaluation.

The RAEC was implemented based on [117]. A conventional, non-robust AEC was also emulated by adjusting the parameters and modifying the RAEC (e.g., removal of ERN, inclusion of a DTD, only one adaptive iteration per block of data, etc.) to provide a basis for the non-robust AEC. A Hamming window with a frame size of 512 and 75% overlap was used for the postfilter. The weighting factor for the DD estimator (66) was $\alpha_{\mathrm{DD}} = 0.98$. The masking threshold was estimated using the "Psychoacoustic Model 2" from the MPEG-1 audio coding standard [63]. The residual echo estimation based on the minimum of two methods, the equivalent transfer function method and the coherence function method [48] (abbreviated as ETF+CF), was implemented as a traditional RES method.

For the AEC performance evaluation, the true echo return loss enhancement (TERLE) (i.e., echo return loss enhancement (ERLE) measured without v) was used. In order to determine how faithfully the RES output represents the near-end signal and how the RES affects the overall system performance, background noise reduction was not performed. Specifically, the AEC output e and the postfiltered AEC output $\hat{v}$ were evaluated, with the near-end signal v treated as the reference containing both the near-end speech and the air conditioner noise. Then for the RES performance evaluation, the segmental signal-to-residual echo ratio (SSRR), the log-spectral distortion (LSD), and the Performance Evaluation of Speech Quality (PESQ) score were chosen. The wide-band mode was used for the PESQ score. The SSRR is defined similarly to the SSNR as

$$\text{SSRR} = \frac{1}{\mathcal{J}} \sum_{m=0}^{\mathcal{J}-1} \mathcal{T}\left\{ 10 \log_{10} \frac{\sum_{n=0}^{K-1} v^2[n + \frac{Km}{4}]}{\sum_{n=0}^{N-1} b^2[n + \frac{Km}{4}]} \right\}, \tag{90}$$

where the true residual echo b is calculated from $b = e - v$.

Table 6 provides the averaged TERLE from the robust and the non-robust AEC. Table 7, 8, and 9 show the averaged SSRR, LSD, and PESQ score, respectively, from the two AEC systems. The better results are reflected by boldface numbers in all tables, and the results from the AEC outputs before the RES are provided as baseline scores in Table 7, 8, and 9. Overall, using the RAEC over the non-robust version increases the TERLE by over 10 dB, whereas the proposed system approach consistently provides better SSNR, LSD, and PESQ when compared to the traditional RES approach. The RAEC without any RES gives better quality measures than the non-robust one with RES. In Table 7, lower input SSNR simply means that the near-end signal power is higher since it contains more air conditioner noise. Thus the baseline SSRR is also higher since the residual echo power is now much smaller compared to the near-end signal power. In Table 8, the postfiltered RAEC output scores worse than the unprocessed one due to the distortion introduced by the suppression gain. The distortion may in fact come from the background

noise suppression. Since our system tends to not suppress the background noise, it introduces less distortion than the traditional RES after postfiltering. In Table 9, the traditional RES may not significantly improve the PESQ score in all cases. On the other hand, our proposed method always improves the score by as much as 0.53 and 0.79 when compared to the unprocessed outputs of the robust and the non-robust AEC, respectively. Based on PESQ, our system combination of the RAEC and the proposed RES delivers the highest overall perceptual quality.

Table 6: TERLE comparison (higher is better).

Input SSNR	0 dB	10 dB	20 dB	30 dB
Conv. AEC	14.69	18.96	19.32	19.48
Robust AEC	**24.88**	**27.01**	**29.64**	**31.21**

Table 7: SSRR comparison (higher is better).

Input SSNR	Conv. AEC	ETF+CF	Proposed	Robust AEC	ETF+CF	Proposed
0 dB	22.55	23.76	**24.78**	29.44	28.94	**29.51**
10 dB	20.26	22.14	**23.34**	25.63	25.31	**26.25**
20 dB	16.02	18.41	**20.71**	22.74	22.33	**24.23**
30 dB	12.43	14.33	**18.19**	18.87	18.24	**21.92**

Table 8: LSD comparison (lower is better).

Input SSNR	Conv. AEC	ETF+CF	Proposed	Robust AEC	ETF+CF	Proposed
0 dB	1.47	1.11	**0.93**	0.34	0.41	**0.35**
10 dB	1.08	0.75	**0.64**	0.34	0.43	**0.35**
20 dB	1.07	0.69	**0.55**	0.28	0.39	**0.31**
30 dB	1.06	0.65	**0.48**	0.24	0.37	**0.27**

Table 9: PESQ comparison (higher is better).

Input SSNR	Conv. AEC	ETF+CF	Proposed	Robust AEC	ETF+CF	Proposed
0 dB	2.01	2.32	**2.60**	3.84	3.79	**4.04**
10 dB	2.25	2.78	**3.04**	3.46	3.58	**3.91**
20 dB	2.41	2.80	**3.04**	3.36	3.52	**3.89**
30 dB	2.75	3.02	**3.16**	3.55	3.58	**3.89**

Figure 7 shows the spectrograms (up to 4 kHz) of the near-end signal, the RAEC output, and the two postfiltered results at 10 dB SSNR. Although ST and DT

information are not used by the RAEC, they are indicated in the figure to show the RES performance under different near-end signal mixing environments. We can see that the proposed method almost completely removes the residual echo. Informal listening tests show that in the traditional RES approach, the residual echo is very weak but still audible. In our proposed RES approach, the residual echo is almost imperceptible. Figure 8 compares the TERLE from the RAEC with the two RES methods at 30 dB SSNR. It shows that our proposed RES achieves higher overall TERLE compared to the traditional RES approach. According to [65], 45 dB TERLE during single talk and 30 dB TERLE during double talk are recommended when no acoustic noise is added. The proposed system achieves more than 45 dB TERLE during single talk and around 30 dB TERLE during double talk when the near-end signal energy is low. The TERLE may be below 30 dB during double talk when the residual echo is already sufficiently masked by the near-end signal.

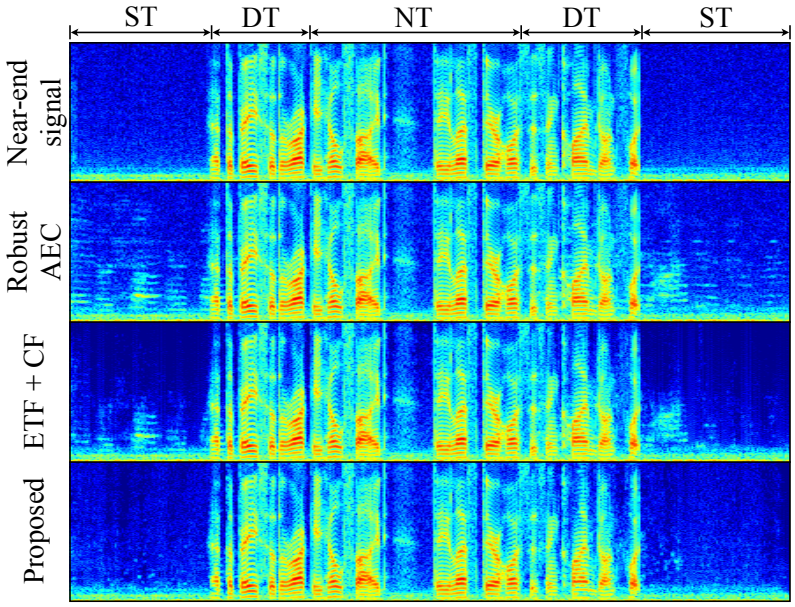


Figure 7: Spectrograms comparing the two RES methods.

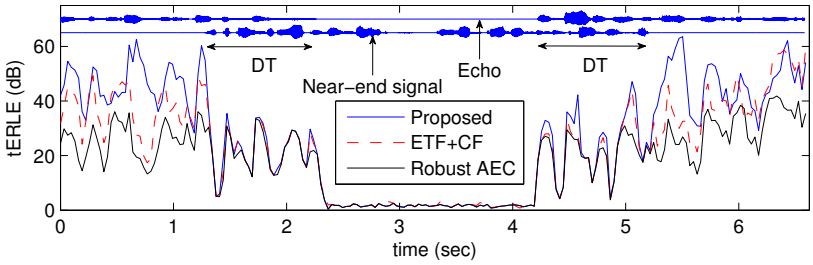


Figure 8: Comparison of TERLE at 30 dB SSNR.

3.3.2 System Approach to Acoustic Echo Cancellation

16 kHz 16-bit PCM recordings of female and male speech signals from the TIMIT database were used for both the far-end and the near-end signals to implement a 2×2 multi-channel AEC system. The far-end signals were normalized to $[-1, 1]$ range, and the echo signals were re-scaled to produce around 15 dB echo return loss before the addition of the near-end speech of equal power. At each end, there were two talkers talking with an overlap of 2 seconds. In addition, the far-end and the near-end talkers took turns talking with an overlap of 2 seconds. Zero-mean white Gaussian noise (WGN) at 40 dB SNR was added to the far-end signals. Air conditioner noise at the SSNR of 0 to 30 dB with 10 dB increments was added to the microphone signals. The measured room responses with $T_{60} = 220$ ms were truncated to 2048 taps (128 ms). 11 test pairs of near-end signal and far-end signal with an average length of 60 seconds were created. To evaluate the steady state performance, the first 10 seconds of each test pair were removed prior to evaluation.

The RAEC component was based on the frequency-domain least mean squares (FDLMS) algorithm with adaptive regularization, ERN, and block-iterative adaptation [117, 119]. The following parameters were used for RAEC: L (filter length) = 2048, B (block size) = $2L$, α (FBLMS step-size) = 0.15, $\gamma = 1$ (regularization parameter), $\beta = 0.998$, and $\eta_1 = \eta_2 = 10$. The baseline RAEC was executed for 8 iterations per block, while each RAEC component in the proposed AEC system was executed for 4 iterations per block. A Hamming window with a frame size of

512 and 75% overlap was used for the postfilter. The weighting factor for the DD estimator (66) was $\alpha_{\mathrm{DD}} = 0.98$.

For the AEC performance evaluation, the TERLE, the SSRR, and the LSD were used. In order to determine how faithfully the AEC output represents the near-end signal, neither RES nor background noise reduction was performed. Specifically, the RAEC output and the proposed AEC system output were evaluated, with the near-end signal v treated as the reference containing both the near-end speech and the air conditioner noise. TERLE was not calculated when the far-end speakers were inactive. On the other hand, SSRR and LSD were evaluated throughout the data regardless of the near-end speakers activity to take into account the performance of both single-talk and double-talk situations. Since our hybrid approach contains nonlinear processing, e.g., the frequency selection (89), SSRR and LSD were used to measure the possible near-end signal distortion due to the hybrid approach.

Table 10, 11, and 12 show the TERLE, SSRR, and LSD, respectively, averaged over the two channels from the RAEC and the proposed AEC system. The better results are reflected by boldface numbers in all tables. Generally, the overall performance increases when the input SSNR increases except for Table 11, where the SSRR decreases as the input SSNR increases. However, higher input SSNR simply means that the near-end signal power is lower since it contains less air conditioner noise and thus the lower SSRR. Overall, the proposed AEC system consistently outperforms the original RAEC. Our proposed system improves the tERLE by about 5 dB and the SSRR by about 4 dB, while lowering the LSD by about half. We can easily observe that the net effect of the proposed system is to further reduce the residual echo of the output of RAEC without adversely affecting or distorting the near-end signal information.

Figure 9 shows the spectrograms of the RAEC output, the proposed AEC system output, the residual echo of the RAEC, and the residual echo of the proposed AEC

Table 10: TERLE comparison (higher is better).

Input SSNR	0 dB	10 dB	20 dB	30 dB
RAEC	17.97	19.76	21.04	21.56
Proposed	**23.25**	**25.74**	**26.53**	**27.19**

Table 11: SSRR comparison (higher is better).

Input SSNR	0 dB	10 dB	20 dB	30 dB
RAEC	26.25	22.33	19.26	16.27
Proposed	**29.86**	**26.63**	**23.27**	**19.92**

Table 12: LSD comparison (lower is better).

Input SSNR	0 dB	10 dB	20 dB	30 dB
RAEC	0.57	0.56	0.51	0.47
Proposed	**0.32**	**0.27**	**0.24**	**0.22**

system at 10 dB SSNR. For clarity, the spectrum of only up to 4 kHz is shown. We can clearly see that the residual echo of the proposed AEC system is greatly reduced compared to that of the RAEC alone especially during double talk. The performance gain during double talk comes from better control of the proposed ERN threshold (85), whereas the gain during single talk comes from the output selection process (89) of the hybrid approach. Figure 10 compares the TERLE of the RAEC output and the proposed AEC system output at 30 dB SSNR. This plot is time-aligned with the spectrogram plot to illustrate the TERLE performance for different near-end mixing environments. While there are improvements also during single talk, we note that the TERLE during double talk can be boosted as much as 10 dB and is almost on a par with the performance during single talk. Overall, the proposed system approach greatly reduces the residual echo of the AEC system, especially during double talk where a traditional AEC system may freeze the adaptation entirely in such a situation. The proposed AEC system produces a near-end signal of higher quality compared to the RAEC alone.

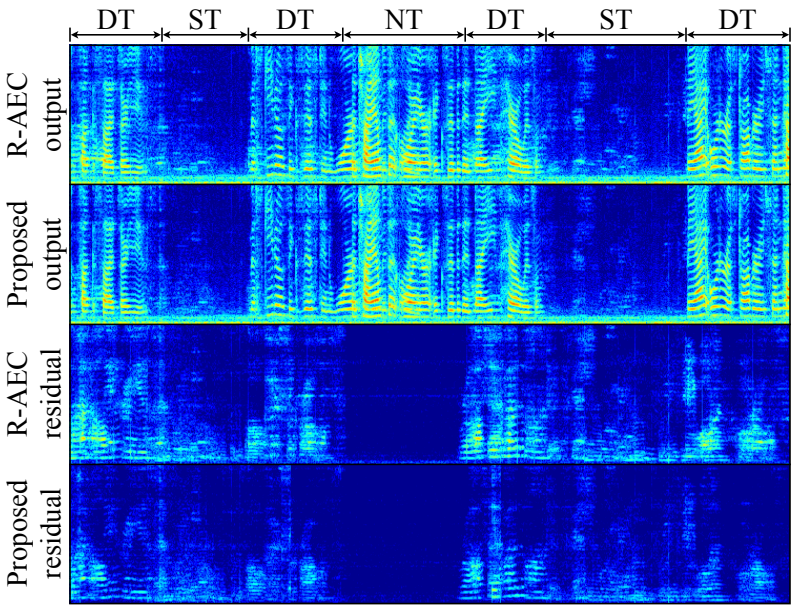


Figure 9: Spectrograms comparing two AEC systems (left channel).

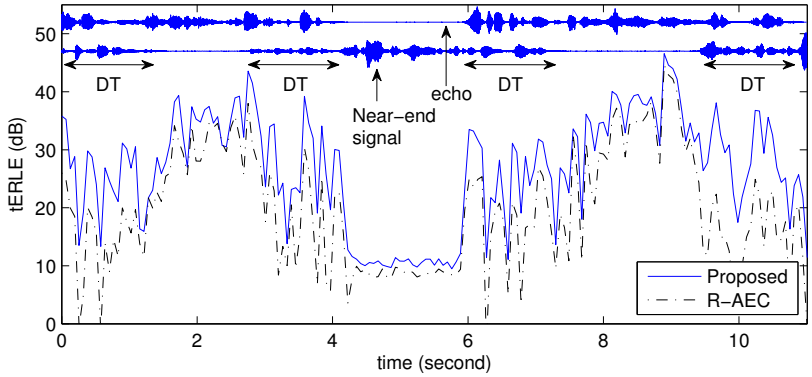


Figure 10: Comparison of TERLE at 30 dB SSNR (left channel).

3.3.3 Application to the *Kinect*™ Audio

The system approach to AEC described in Section 3.2 has been successfully implemented and tested for the *Kinect*™ audio pipeline.[1] A block diagram of the *Kinect*™ audio pipeline is shown in Figure 11, which consists of a four-microphone linear array, a constant tone removal (CTR) unit, a multi-channel acoustic echo cancellation (MCAEC) unit, a beamformer (BF), a RES unit, and a noise suppression (NS) unit. The CTR block is used to remove the fan noise inherent to the *Kinect*™ device.

[1] Work done at Microsoft Research as an research intern in the summer of 2012.

The MCAEC cancels the echo generated by the surround sound loudspeakers. The BF combines all four outputs after the MCAEC, and the RES and the NS further enhance the signal by suppressing the residual echo and the background noise.

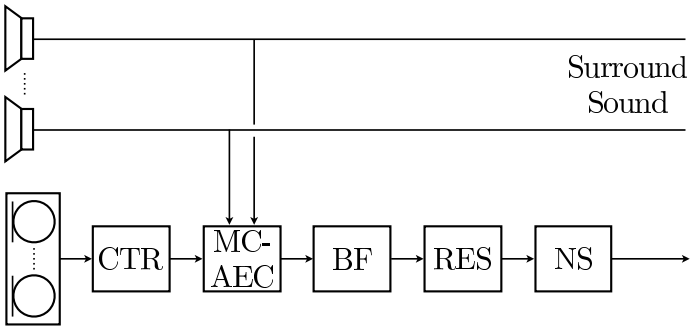


Figure 11: A block diagram of the *Kinect*™ audio pipeline.

Since a typical home theatre setup has five channels and the *Kinect*™ microphone array has four microphones, the MCAEC needs to keep track of 20 echo paths and is prone to the non-uniqueness problem. To reduce the complexity of the MCAEC and save computation, a dual-layered AEC structure shown in Figure 12 [108] was used. The idea is to first identify each echo path by playing a calibration tone when the user first set up the system, assuming that the loudspeaker and the *Kinect*™ microphone array positions will stay intact after the initial setup. Therefore, for each echo path a fixed filter needs to be calculated, and for each microphone an echo signal from all the loudspeakers needs to be estimated by using the fixed filters. The estimated echo signal from the fixed filters serve as the reference signal for the microphone, and echo cancellation is performed using only one adaptive filter for each microphone. This structure alleviates the non-uniqueness problem and saves five times the computational load in our estimate. However, whenever the loudspeaker or the microphone array positions change, the whole system needs to be recalibrated.

Since the whole system contains multiple blocks that contain a total of 180+ parameters across the entire pipeline, tuning of the system becomes tricky and time-consuming if done by hand. Therefore, an automated tuning scheme was used to

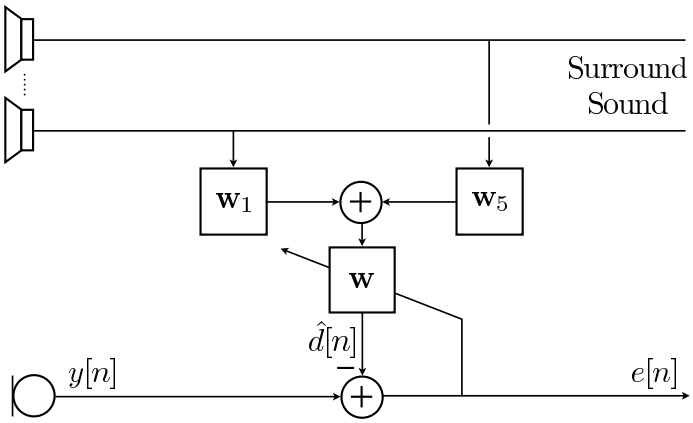


Figure 12: A block diagram of the dual-layered AEC. Only one of the four microphones is shown.

optimize for the parameters of the system. A data corpus is created to encompass the general user scenarios, with various loudspeaker levels, various clean speech levels, different type of loudspeaker signals, and various user position relative to the microphone array. The loudspeaker levels range from 65 to 75 dB(C) sound pressure level (SPL) while the clean speech levels range from 60 to 65 dB(C) SPL. Several loudspeaker signals such as game sound, stereo music, or movie surround sounds were used. The clean speech contains either long utterances from the Harvard speech database or some short Xbox commands that is custom made. The users are located at 1 to 4 meters away from the microphone array with a 1-meter increment and can be standing at either the left, the center, or the right of the array with a 1-meter increment. Therefore there are totally 12 possibile speaker locations. The users are mostly at fixed locations, but some are moving in order to tune the tracking performance of the beamformer.

A Gaussian minimization algorithm [108] was used to optimize the whole system. The optimization procedure was done from the first block to the last block, where each block is tuned one at a time. Within each block the parameters are tuned one at a time while fixing all the other parameters. Several iterations may be performed when tuning each block as well as going from block to block. A composite

optimization criterion (similar to [111]) was used, where objective measures, such as PESQ or ERLE, were used to optimize for the speech quality while the word error rate (WER) and signal-to-error ratio (SER) were used to optimize for the automatic speech recognition (ASR) system performance. The reason for using the composite optimization criterion was that tuning the system based solely on one measure alone may skew the parameters in favor of that particular measure at the expense of degrading the performance of other measures. While a composite optimization criterion using speech quality measures (PESQ or ERLE) alone may be sufficient for human listeners, the system may favor only speech quality improvement at the expense of degrading the ASR performance. Therefore, a combination of both the speech quality measures and the speech recognition performance measures guarantees that the system performs well for both human listeners and ASR systems.

The system has to function mostly in continuous double talk scenarios, e.g., when a user is issuing a command during either movie, music or game sound playback. The original AEC in the *Kinect*™ audio uses the traditional least mean squares (LMS) adaptive filter with a DTD to freeze the update during double talk. By integrating the system approach to AEC, the echo cancellation performance can be greatly improved during such continuous double talk scenario. Table 13 shows the simulation results with the baseline system using the DTD-based traditional LMS adaptive filter and the results with the system approach to AEC.[2] For convenience of the discussion, the letter "A" represents different AEC schemes, and "P" represents post-AEC, i.e., applying a multi-channel AEC after the beamformer to further reduce the echo. "A1" represents the system approach to AEC operating in multi-channel configuration without the calibration procedure, while "A2" utilizes both the calibration and the system approach to AEC. "P1" indicates the the post-AEC is applied while "P2" indicates that it is turned off. The post-AEC can be seen as an extra procedure to

[2]Only relative improvement compared to the baseline system performance is shown.

extract the echo signal that is not correctly estimated by the calibration procedure, due to the measurement noise during the calibration or changes in the room condition after the calibration.

Table 13: Simulation results. FFTs indicate the increase in computational complexity in terms of the number of FFT operations. The system approach to AEC inevitably increases the computational cost due to the two-pass adaptation.

method	PESQ	ERLE (dB)	SER	WER	FFTs
baseline	–	–	–	–	–
A1P0	+0.16	−1.2	−45.5%	−25.3%	+78.8%
A1P1	+0.05	−1.0	−28.7%	−1.7%	+105%
A2P0	+0.20	+7.8	−35.7%	−19.6%	+ 9.9%
A2P1	**+0.24**	**+9.6**	**−63.0%**	**−60.4%**	+35.4%

From the simulation results we observe that both the perceptual quality (PESQ and ERLE) and the ASR performance (SER and WER) are greatly enhanced with "A2"; the best performance was achieved through "A2P1". We note that the overall ERLE of "A1" is slightly degraded, likely due to the non-uniqueness problem that slows down the convergence speed of the multi-channel AEC without a decorrelation procedure. Furthermore, the computational cost of "A1" is greatly increased due to the increase in the number of echo paths that needs to be tracked by the multi-channel AEC. Nonetheless, the ASR performance is still improved, due to the robustness of the system approach during double talk.

CHAPTER IV

DECORRELATION BY SUB-BAND RESAMPLING

As we have discussed in Section 2.2.1, the non-uniqueness problem arises during multi-channel acoustic echo cancellation (MCAEC) due to the correlation between reference signals (i.e., far-end multi-channel microphone signals) that degrades the convergence performance of adaptive filtering algorithms [104]. The MCAEC solution (to prevent return of the far-end signal at the near-end) is in fact dependent on the far-end room impulse response and must reconverge, for example, when the far-end speech activities change such as changing the talker position or even the talker all together. Applying a decorrelation procedure before near-end playback can improve the tracking of the signal responses along the echo paths with, hopefully, a minimal side effect on both the audio quality and the original signal statistics [118]. Key criteria for an ideal decorrelation procedure are as follows:

- retains the original audio quality and sound image of the far-end.
- retains the original excitation characteristics of the echo paths.
- retains the original signal statistics used for the adaptive filters.
- scalable to a large number of channels.
- requires low computational complexity.

A handful of inter-channel decorrelation procedures have been proposed in the past to alleviate the non-uniqueness problem and the associated tracking issue, e.g., [7, 34, 55, 89, 106, 107]. However, these decorrelation techniques may not achieve an optimal steady-state performance of an adaptive filter and are usually performed in

a full-band manner, leaving no possibility for "frequency-selective" decorrelation. By frequency-selectivity we mean performing varying degrees of decorrelation for different frequency channels. Although the phase modulation procedure [55] allows a perceptually motivated frequency-selective choice of phase modulation parameters by employing sub-band decomposition, the quantitative relationship between decorrelation and phase modulation remains unclear at least analytically. Moreover, the resultant audio quality is traditionally examined in an *ex post facto* manner rather than an active factor in the design of the decorrelation procedure.

Recently, Wada and Juang proposed a decorrelation algorithm by a resampling procedure [118] which was motivated by the analysis of the sampling rate mismatch problem inherent to audio processing using distributed audio devices [98]. The resampling procedure was extended to the frequency domain [120], the time domain [123], and the sub-band resampling (SBR) [124]. When applied to our noise-robust frequency-domain MCAEC system [117,119,125], decorrelation by resampling results in a faster echo path tracking than other decorrelation procedures while keeping a minimal distortion to the signal quality and statistics [118, 120]. The power of SBR is that the amount of decorrelation can be finely controlled for a better perceptual quality, i.e., the amount of resampling can be arbitrarily controlled in each frequency bin such that the perceptually less significant sub-bands can be more aggressively decorrelated while still preserving a high speech quality [124, 125, 129].

We have experimentally demonstrated in [118, 120] the effectiveness of decorrelation by resampling on MCAEC and in [124, 125, 129] the superior processed speech quality of SBR. The objective of this section is to extend these contributions, provide a deep analysis of the performance bounds of the resampling procedure, and devise a proper overall scheme for SBR. Specifically, we analyze the links between the proposed resampling, the level of decorrelation, and the achievable steady-state misalignment

performance. We then derive new closed-form expressions to demonstrate how resampling affects the misalignment for different types of reference signals, i.e., a white Gaussian noise or a speech signal, with or without the far-end room impulse response. Following these analyses, we provide a novel, theoretically justifiable, and perceptually motivated SBR strategy for achieving fast MCAEC convergence with a minimal signal distortion. In our experimental evaluation, we show that the proposed SBR scheme outperforms other decorrelation methods in terms of the convergence rate and the processed speech quality.

4.1 Decorrelation by Resampling

4.1.1 Misalignment Problem Revisited

It can be shown [5, 34, 75] that the steady-state misalignment for a two-channel frequency-domain adaptive filter (FDAF) after convergence can be approximated by

$$\zeta \approx 10\log_{10}\left(\frac{1-\lambda}{2}\frac{\sigma_v^2}{\|\mathbf{h}\|_2^2}\mathrm{tr}\{\mathbf{S}^{-1}\}\right) \quad \text{(dB)}, \tag{91}$$

where $0 \ll \lambda < 1$ is a forgetting factor, σ_v^2 is the variance of the near-end noise,[1] $\mathbf{S}$ is the spectral density matrix of the reference signals, and $\mathrm{tr}\{\cdot\}$ is the trace operator. The spectral density matrix and its inverse are given by [75]

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \tag{92}$$

$$\mathbf{S}^{-1} = \begin{bmatrix} \mathbf{S}_1^{-1} & \mathbf{0}_{2L\times 2L} \\ \mathbf{0}_{2L\times 2L} & \mathbf{S}_2^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{2L\times 2L} & -\mathbf{S}_{12}\mathbf{S}_{22}^{-1} \\ -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & \mathbf{I}_{2L\times 2L} \end{bmatrix}, \tag{93}$$

[1] Without the near-end noise but with properly decorrelated reference signals, an adaptive algorithm converges to the true solution. This is reflected in (91) since ζ goes to negative infinity as σ_v^2 goes to zero, i.e., the near-end noise sets the lower bound of (91).

and the submatrices are given by

$$\mathbf{S}_1 = \mathbf{S}_{11}(\mathbf{I}_{2L\times 2L} - \mathbf{C}_{12}) \tag{94}$$

$$\mathbf{S}_2 = \mathbf{S}_{22}(\mathbf{I}_{2L\times 2L} - \mathbf{C}_{12}) \tag{95}$$

$$\mathbf{S}_{ij} = \mathrm{diag}\{S_{x_i x_j}[0], \ldots, S_{x_i x_j}[2L-1]\}, \quad i,j = 1,2 \tag{96}$$

$$\mathbf{C}_{12} = \mathrm{diag}\{C_{x_1 x_2}[0], \ldots, C_{x_1 x_2}[2L-1]\}, \tag{97}$$

where $S_{x_p x_p}[k]$, $p = 1, 2$, $S_{x_i x_j}[k]$, $i \neq j$, and $C_{x_1 x_2}[k]$, are respectively the power spectral density (PSD), the cross-spectral density (CSD), and the coherence (or magnitude-squared coherence) function [16] between the two signals in the k^{th} frequency bin. The operator $\mathrm{diag}\{\cdot\}$ forms a diagonal matrix. Note that a frame length of $2L$ is used so that our analysis in later sections will be consistent with the overlap-save (OLS) FDAF structure [102].

Let $r_{x_i x_j}[n]$ be the auto- and cross-correlation function for $i = j$ and $i \neq j$, respectively. Depending on whether $i = j$ or $i \neq j$, the PSD or the CSD is given by

$$S_{x_i x_j}[k] = \sum_{n=-\infty}^{\infty} r_{x_i x_j}[n] e^{-j\frac{\pi}{L}kn}. \tag{98}$$

The coherence, which is a real-valued function that represents the amount of correlation between two signals in the frequency domain, is defined in terms of the PSDs and the CSD as

$$C_{x_1 x_2}[k] \equiv \frac{\left|S_{x_1 x_2}[k]\right|^2}{S_{x_1 x_1}[k] S_{x_2 x_2}[k]}, \quad 0 \leq C_{x_1 x_2}[k] \leq 1. \tag{99}$$

Using (92), the trace of $\mathbf{S}^{-1}$ is expressed in terms of the PSDs and the coherence as

$$\mathrm{tr}\{\mathbf{S}^{-1}\} = \sum_{k=0}^{2L-1} (1 - C_{x_1 x_2}[k])^{-1} (S_{x_1 x_1}^{-1}[k] + S_{x_2 x_2}^{-1}[k]). \tag{100}$$

First note that for the trivial case of $x_1[n] = x_2[n]$, the coherence becomes one, the spectral density matrix $\mathbf{S}$ becomes singular, $\mathbf{S}^{-1}$ does not exist, and the misalignment is unbounded. On the other hand, when there is only one source at the far-end,

the highly correlated reference signals result in a spectral density matrix that is still close to singular. Consequently, the misalignment is negatively affected by the highly correlated reference signals as dictated by (91) and (100). A decorrelation procedure is required to lower the coherence between the reference signal channels and consequently improve the misalignment of MCAEC.

4.1.2 Proposed Resampling Method

Our resampling approach was first introduced in [118,120] to address the non-uniqueness problem in MCAEC. The idea behind decorrelation by resampling is to exploit the effect of sampling rate mismatch examined in [98], where a slight sampling rate mismatch between audio channels of a few hundred parts per million ($\sim 0.01\%$) is enough to break down the correlation structure necessary for a sufficient MCAEC performance. Conversely, by artificially introducing a sampling rate mismatch between the highly correlated reference signals, we should achieve improved MCAEC performance while minimizing the distortion to the reference signal quality.

For discrete-time signals, decorrelation by time expansion/compression is implemented by up/downsampling a signal to a different sampling rate f_s' and playing back the resampled signal at the original rate f_s, where the expansion/compression ratio is related to the resampling ratio by $R = f_s'/f_s$. By properly adjusting the resampling ratios across channels, a variable delay is introduced across channels to decorrelate the reference signals.

Figure 13 shows the proposed resampling scheme that was successfully used in [120, 123–125, 128, 129]. The variable delay between the two channels is achieved by properly resampling the reference signals, and the delay is varied smoothly (mainly without disruptive discontinuity) across time to eliminate the potential distortion associated with sudden changes in the delay. Furthermore, it is shown in [124,125,129] that this resampling procedure with SBR achieves a higher audio quality compared to

the nonlinearity proposed in [7,34,89] or the phase modulation in [55] for an equivalent degree of decorrelation as measured by the coherence.

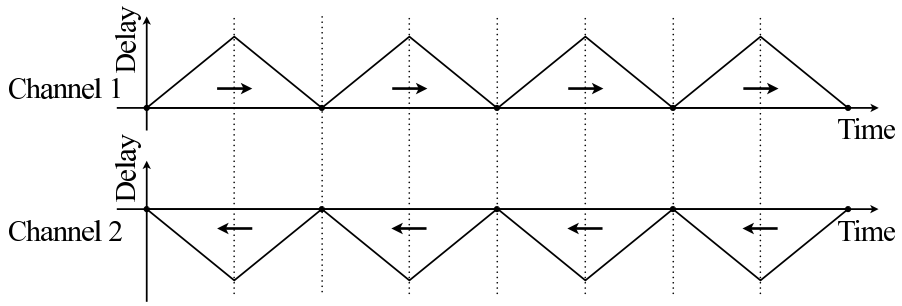


Figure 13: The proposed resampling scheme that achieves variable delay across the two reference signal channels for SAEC. The dotted lines represent signal blocks of length N and the arrows represent the direction of the signal shift after resampling. We hereby reserve the term "block" with length N for the resampling process and "frame" for the OLS FDAF structure. Refer to Section 4.3 for more details.

4.2 Coherence/Misalignment vs. Resampling

In this section, we focus on the resampling procedure and derive the relationships between resampling, the level of decorrelation, and the misalignment of SAEC. Once the links are established, we can properly adjust the resampling ratios in each frequency bin to achieve an improved convergence rate with minimal signal degradation. Since (91) and (100) already show the relationship between the coherence and the misalignment, we need to find out how resampling influences the coherence.

The resampling process for a discrete-time signal is related to the time-scaling process for a continuous-time signal. By time expanding a continuous-time signal $x(t)$ to $x(t/R)$ with an expansion ratio $R > 1$, the delay is steadily built up over time between the original signal and the time-expanded signal. Intuitively, the coherence between $x(t)$ and $x(t/R)$ should decrease due to the delay buildup. Similarly, the time-compression process can be easily obtained from $x(t/R') = x(Rt)$ by choosing a compression ratio $R' = 1/R$.

This time-scaling process and its effect on the coherence have been studied in [124].

However, a similar relationship for discrete-time signals has yet to be established. Here we analyze the resampling procedure for both continuous-time and discrete-time signals and show how the coherence is altered. We then study the proposed resampling scheme in Figure 13 and show how it effectively decorrelates the reference signals and improves the misalignment of SAEC.

4.2.1 Link between Coherence and Continuous-Time Scaling

Given two wide-sense-stationary random processes x_t and y_t, the coherence at each frequency ω is measured by

$$C_{xy}(\omega) \equiv \frac{|S_{xy}(\omega)|^2}{S_{xx}(\omega)S_{yy}(\omega)}, \quad 0 \leq C_{xy}(\omega) \leq 1, \tag{101}$$

with $C_{xy} = 1$ being perfectly correlated and $C_{xy} = 0$ being uncorrelated. The CSD $S_{xy}(\omega)$ is given by

$$S_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(\tau)e^{-j\omega\tau}\,\mathrm{d}\tau \equiv \mathcal{F}\{R_{xy}(\tau)\}, \tag{102}$$

where $\mathcal{F}\{\,\cdot\,\}$ is the continuous-time Fourier transform (CTFT) and $R_{xy}(\tau) = \mathrm{E}\{x_t y^*_{t-\tau}\}$ is the cross-correlation. $S_{xx}(\omega)$ and $S_{yy}(\omega)$ are PSDs of x and y, respectively, and are calculated by $\mathcal{F}\{R_{xx}(\tau)\}$ and $\mathcal{F}\{R_{yy}(\tau)\}$.

For $x(t)$ and $y(t)$ as the actual realizations of the stochastic processes in continuous time, the cross-correlation between the two signals is estimated by

$$\tilde{R}_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y^*(t-\tau)\,\mathrm{d}t, \tag{103}$$

and the CSD is given by $\tilde{S}_{xy}(\omega) = \mathcal{F}\{\tilde{R}_{xy}(\tau)\} = X(\omega)Y^*(\omega)$, where $X(\omega) = \mathcal{F}\{x(t)\}$ and $Y(\omega) = \mathcal{F}\{y(t)\}$. The PSDs of $x(t)$ and $y(t)$ are given by $\tilde{S}_{xx}(\omega) = |X(\omega)|^2$ and $\tilde{S}_{yy}(\omega) = |Y(\omega)|^2$, respectively. However, the coherence in this case is equal to one for (101) since only the instantaneous realizations are used for calculation without taking into account the mathematical expectation. Therefore, the CSD is estimated in practice by averaging over short-time evaluations. That is, let $w(t)$ be a window

function with the support $t \in [0, T]$, $w_m(t) = w(t - mt_0)$ be the m^{th} window with a delay of mt_0, where $m = 0, 1, \ldots, M-1$, $t_0 \leq T$, and M is the number of signal blocks. Then the CSD is estimated by [122]

$$\hat{S}_{xy}(\omega) = \frac{1}{M} \sum_{m=0}^{M-1} X_m(\omega) Y_m^*(\omega), \tag{104}$$

where $X_m(\omega) = \mathcal{F}\{x(t)w_m(t)\}$ and $Y_m(\omega) = \mathcal{F}\{y(t)w_m(t)\}$. The PSD can be similarly estimated. The coherence is estimated as

$$\hat{C}_{xy}(\omega) \equiv \frac{\left| \sum_{m=0}^{M-1} X_m(\omega) Y_m^*(\omega) \right|^2}{\left(\sum_{m=0}^{M-1} |X_m(\omega)|^2 \right) \left(\sum_{m=0}^{M-1} |Y_m(\omega)|^2 \right)}. \tag{105}$$

By time expanding a continuous-time signal $x(t)$ to $x(t/R)$ with an expansion ratio $R > 1$, the delay is steadily built up over time between the original signal and the time-expanded signal. Intuitively, the cross-correlation between $x(t)$ and $x(t/R)$ should go down due to the delay buildup. We can quantify this effect through the analysis below, which can be similarly applied to time compression $x(t/R') = x(Rt)$ by choosing a compression ratio $R' = 1/R$. From now on we will always assume $R > 1$.

Let $x(t) = e^{j\omega_0 t}$, $y(t) = x(t/R)$, and $w(t)$ be the rectangular window that is zero outside $t \in [0, T]$. The CTFTs of the signals and the m^{th} window function are

$$X(\omega) = 2\pi\delta(\omega - \omega_0), \tag{106}$$

$$Y(\omega) = 2\pi R\delta(\omega - \omega_0/R), \tag{107}$$

$$W_m(\omega) = T \operatorname{sinc}(\omega T/2) e^{-j\omega(mt_0 + T/2)} \tag{108}$$

where $\delta(x)$ is the Dirac delta function and $\operatorname{sinc}(x) \equiv \sin(x)/x$. Using the convolution theorem $\mathcal{F}\{x(t)w_m(t)\} = \frac{1}{2\pi} X(\omega) * W_m(\omega)$, where $*$ is the convolution operator, the CTFTs of the windowed signals $x_m(t)$ and $y_m(t)$ are given by

$$X_m(\omega) = T \operatorname{sinc}\big((\omega - \omega_0)\tfrac{T}{2}\big) e^{-j(\omega - \omega_0)(mt_0 + \frac{T}{2})}, \tag{109}$$

$$Y_m(\omega) = RT \operatorname{sinc}\big((\omega - \tfrac{\omega_0}{R})\tfrac{T}{2}\big) e^{-j(\omega - \frac{\omega_0}{R})(mt_0 + \frac{T}{2})}, \tag{110}$$

and the frequency contents at ω_0 are given by

$$X_m(\omega_0) = T, \tag{111}$$

$$Y_m(\omega_0) = RT\,\mathrm{sinc}\big(\tfrac{\Delta R\omega_0}{R}\tfrac{T}{2}\big)e^{-j\frac{\Delta R\omega_0}{R}(mt_0+\frac{T}{2})} = Ae^{-j\frac{\Delta R}{R}\omega_0 t_0 m}, \tag{112}$$

where A is a complex constant independent of m and $\Delta R \equiv R - 1$. Using (105), the coherence estimate at ω_0 is

$$\hat{C}_{xy}(\omega_0) = \frac{\big|\sum_{m=0}^{M-1} TA^* e^{j\frac{\Delta R}{R}\omega_0 t_0 m}\big|^2}{\big(\sum_{m=0}^{M-1} T^2\big)\big(\sum_{m=0}^{M-1}|A|^2\big)} = \left[\frac{1}{M}\frac{\sin\big(\frac{\Delta R}{2R}\omega_0 t_0 M\big)}{\sin\big(\frac{\Delta R}{2R}\omega_0 t_0\big)}\right]^2. \tag{113}$$

First of all, we note that (113) is independent of the window size T, which only contributes as a constant factor, and the phase term goes away after taking the absolute value. Second, if $M = 1$, (113) is always equal to one since it is calculated over only a single instance. Third, (113) is always one also if $\Delta R = 0$ since there is no time scaling. Finally, for $M > 1$ and $\Delta R \neq 1$, we can evaluate the reduction in the coherence by the following numerical example.

Suppose the continuous-time signal is bandlimited to $f_c = 8$ kHz at a sampling rate of $f_s = 16$ kHz. If the coherence measurement block size is $N = 2048$ samples that is divided into $M = 8$ sub-blocks with 50% overlap, then the block shift in continuous time becomes $t_0 = \frac{N}{M}\frac{1}{f_s} = 16$ ms. We can fix ΔR at certain values and sweep the signal frequency $f_0 = \omega_0/2\pi \in [0, f_c]$ kHz. By doing so with (113) and selecting $\Delta R = 0.0004$, 0.0008, 0.0012, and 0.0016, we obtain the coherence-frequency plot in Figure 14.

We observe from the plot that for a given ΔR, the coherence is generally inversely dependent on the signal frequency. In particular, we see that before the coherence reaches the first zero, the coherence reduction vs. frequency relationship is approximately linear. Furthermore, for a fixed frequency before the coherence first reaches zero, e.g., $f_0 = 3$ kHz, the coherence also decreases roughly linearly as ΔR increases. Thus (113) provides a way to measure the amount of decorrelation at each frequency

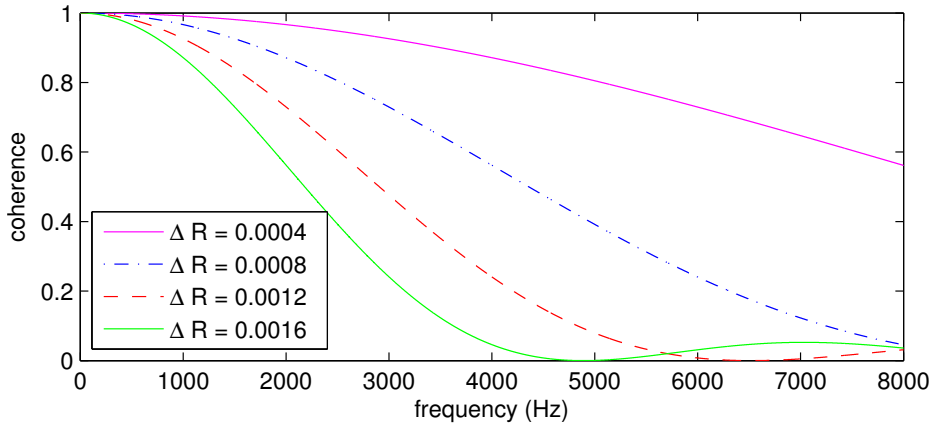


Figure 14: Coherence-frequency plot obtained from (113).

point for a certain expansion ratio R. Conversely, it allows us to control R for a desired amount of decorrelation in terms of the coherence at certain frequency points, e.g., to minimize the distortion of a signal at low frequencies.

4.2.2 Coherence vs. Resampling

The discrete short-time Fourier transform (STFT) of a real-valued discrete-time signal $x[n]$ is given by

$$X_m[l] = e^{-j\omega_l m N_s} \sum_{n=0}^{N-1} x[n + mN_s] w_N[n] e^{-j\omega_l n}, \tag{114}$$

where $X_m[l]$ is the l^{th} discrete Fourier transform (DFT) coefficient in the m^{th} block, N is the block size, $N_s \leq N$ is the block shift size, $\omega_l = 2\pi l/N$, and $w_N[n]$ is a window function that is zero outside $n = 0, \ldots, N-1$. Let $N_s = N$, $x_m[n] = x[n+mN]w_N[n]$, and $w_N[n]$ be the rectangular window, which is chosen to be consistent with the analysis of the OLS FDAF structure in later sections. The discrete STFT of $x_m[n]$ becomes

$$X_m[l] = \sum_{n=0}^{N-1} x_m[n] e^{-j\omega_l n} \equiv \text{DFT}\{x_m[n]\}, \tag{115}$$

and $x_m[n]$ can be expressed in terms of the inverse discrete Fourier transform (IDFT) as (assuming N is divisible by 2 for simplicity)

$$\begin{aligned} x_m[n] &= \text{IDFT}\{X_m[l]\} \equiv \frac{1}{N}\sum_{l=0}^{N-1} X_m[l]e^{j\omega_l n} \\ &= \frac{X_m[0] + X_m[\frac{N}{2}]e^{j\pi n}}{N} + \frac{2}{N}\sum_{l=1}^{\frac{N}{2}-1} \Re\{X_m[l]e^{j\omega_l n}\}, \end{aligned} \tag{116}$$

where $\Re\{\cdot\}$ denotes the real part of a complex number. Since $x_m[n]$ is assumed to be real, only $l = 0, \ldots, \frac{N}{2}$ of conjugate symmetric $X_m[l]$ are required for the reconstruction of $x_m[n]$.

For the signal block $x_m[n]$ of length N, the time-shifting property of the DFT is given by

$$x_m[n - n_0] = \text{IDFT}\{X_m[l]e^{-j\omega_l n_0}\}, \tag{117}$$

where n_0 is a fixed delay. Given a resampling ratio $R > 1$, the delay between the resampled signal and the original signal is $D = RN - N = \Delta RN$, where $\Delta R \equiv R - 1$. Therefore, the sub-sample delay of each sample after resampling is given by

$$d(n) = \frac{D}{RN}n = \frac{\Delta R}{R}n. \tag{118}$$

As the delay varies with respect to the original time after resampling, (118) is used to represent the delay instead of a fixed delay in (117). Resampling by time shifting can be performed by calculating the DFT coefficients $X_m[l]$ of $x_m[n]$, multiplying $X_m[l]$ by the phase-shift term $e^{-j\omega_l d(n)}$, and applying the IDFT on the modified DFT coefficients. This results in the resampled signal $\tilde{x}_m[n]$ given by

$$\begin{aligned} \tilde{x}_m[n] &= x_m[n - d(n)] = \text{IDFT}\{X_m[l]e^{-j\omega_l n\frac{\Delta R}{R}}\} \\ &= \frac{X_m[0] + X_m[\frac{N}{2}]e^{j\frac{\pi n}{R}}}{N} + \frac{2}{N}\sum_{l=1}^{\frac{N}{2}-1} \Re\{X_m[l]e^{j\frac{\omega_l}{R}n}\}. \end{aligned} \tag{119}$$

Note that the time-shifting property assumes the signal $x_m[n]$ to be N-periodic, which is not always true for a general signal. The accuracy of the resampled signal may

decrease as the sample index n gets close to N. Special treatment is needed for block-wise resampling and will be discussed in Section 4.3.

Comparing (119) to (116), we can view $\tilde{x}_m[n]$ as obtained from applying what is referred to as the modified inverse discrete Fourier transform (MIDFT)

$$\tilde{x}_m[n] = \text{MIDFT}\{X_m[l]\} \equiv \frac{1}{N} \sum_{l=-\frac{N}{2}}^{\frac{N}{2}-1} X_m[l] e^{j\frac{\omega_l}{R}n}, \tag{120}$$

where $X_m[-l] = X_m^*[l]$, $l = 0, \ldots, \frac{N}{2}$ (conjugate symmetry is assumed for other values hereafter). From here on, the negative frequency indices, whenever appropriate, will be used for a compact representation of the MIDFT in (120), since the phase term $e^{j\frac{\omega_l}{R}}$ is scaled by R and is no longer sampled at equally-spaced intervals on a unit circle. To account for possible frequency-domain aliasing when a compression ratio $R' = 1/R$ is used and $|\omega_l/R'| > \pi$, $X_m[l]$ has to be first low-pass filtered. This is achieved by simply setting $X_m[l] = 0$, $|l| > \frac{N}{2R}$, and is automatically assumed hereafter for the time compression operation. Using (120), the DFT coefficients of $\tilde{x}_m[n]$ is given by

$$\tilde{X}_m[k] = \frac{1}{N} \sum_{l=-\frac{N}{2}}^{\frac{N}{2}-1} X_m[l] \sum_{n=0}^{N-1} e^{j(\frac{\omega_l}{R}-\omega_k)n} = \frac{1}{N} \sum_{l=-\frac{N}{2}}^{\frac{N}{2}-1} X_m[l] \phi_N\big(\tfrac{l}{R} - k\big), \tag{121}$$

where $k = 0, \ldots, N-1$ and

$$\phi_N(x) \equiv \begin{cases} N, & x = 0 \\ \frac{\sin(\pi x)}{\sin(\frac{1}{N}\pi x)} e^{j\frac{N-1}{N}\pi x}, & \text{otherwise.} \end{cases} \tag{122}$$

Let $x[n]$ be a white Gaussian noise (WGN) process with zero mean and variance σ_x^2. The CSD estimate of $x[n]$ and $\tilde{x}[n]$ is expressed in terms of the ensemble average as

$$\hat{S}_{x\tilde{x}}[k] = \text{E}\{X_m[k]\tilde{X}_m^*[k]\}, \tag{123}$$

and the estimated coherence is given by

$$\hat{C}_{x\tilde{x}}[k] = \frac{\big|\hat{S}_{x\tilde{x}}[k]\big|^2}{\hat{S}_{xx}[k]\hat{S}_{\tilde{x}\tilde{x}}[k]}. \tag{124}$$

We note that $\mathrm{E}\{x_m[n_1]x_m^*[n_2]\} = \sigma_x^2\delta(n_2 - n_1)$ and

$$\begin{aligned}\mathrm{E}\{X_m[l_1]X_m^*[l_2]\} &= \sum_{n_1,n_2} \mathrm{E}\{x_m[n_1]x_m^*[n_2]\}e^{j(\omega_{l_2}n_2 - \omega_{l_1}n_1)} \\ &= \sum_{n=0}^{N-1} \sigma_x^2 e^{j\frac{2\pi}{N}(l_2-l_1)n} = N\sigma_x^2\delta(l_2 - l_1),\end{aligned} \tag{125}$$

where $\delta(\,\cdot\,)$ is the Dirac delta function. Therefore, the first denominator term in (124) is given by

$$\hat{S}_{xx}[k] = \mathrm{E}\{X_m[k]X_m^*[k]\} = N\sigma_x^2. \tag{126}$$

Using (121), the second denominator term in (124) becomes

$$\hat{S}_{\tilde{x}\tilde{x}}[k] = \frac{1}{N^2}\sum_{l_1,l_2} \phi_N\big(\tfrac{l_1}{R} - k\big)\phi_N^*\big(\tfrac{l_2}{R} - k\big)\mathrm{E}\{X_m[l_1]X_m^*[l_2]\} = \frac{\sigma_x^2}{N}\sum_{l=-\frac{N}{2}}^{\frac{N}{2}-1} \big|\phi_N\big(\tfrac{l}{R} - k\big)\big|^2. \tag{127}$$

The numerator term in (124) is given by

$$\big|\hat{S}_{x\tilde{x}}[k]\big|^2 = \left|\frac{1}{N}\sum_{l=-\frac{N}{2}}^{\frac{N}{2}-1} \phi_N^*\big(\tfrac{l}{R} - k\big)\mathrm{E}\{X_m[k]X_m^*[l]\}\right|^2 = \sigma_x^4\big|\phi_N\big(\tfrac{\Delta R}{R}k\big)\big|^2. \tag{128}$$

Substituting the terms in (124) by (126), (127), and (128), we obtain the coherence estimate as

$$\hat{C}_{x\tilde{x}}[k] = \frac{\big|\phi_N\big(\frac{\Delta R}{R}k\big)\big|^2}{\sum_{l=-\frac{N}{2}}^{\frac{N}{2}-1}\big|\phi_N\big(\frac{l}{R} - k\big)\big|^2}, \tag{129}$$

where $k = 0, \ldots, N-1$.

Figure 15 shows the actual measured coherence of a WGN before and after resampling, where only one of the channels in Figure 13 was applied to the WGN and the theoretical coherence is shown as black solid lines. The sampling frequency was 16 kHz, the resampling block size was $N = 2048$ samples, and the coherence was measured across 100 non-overlapping blocks. The resampling ratios $R = 1.0004$, 1.0008, 1.0012, and 1.0016 were used.

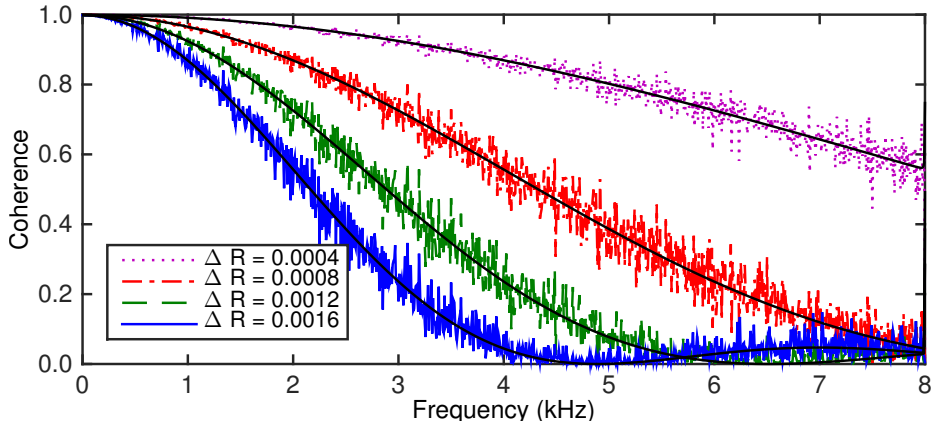


Figure 15: Coherence-frequency plot calculated from a WGN before and after resampling with various ΔR. The black solid curves are calculated from (129). Notice its similarity to Figure 14.

Note that (129) becomes identical to (113) for the same measurement/resampling block size and resampling ratio, albeit with different assumptions of the underlying signals. This WGN model is more useful than the simple sinusoid model for the analysis of the proposed resampling scheme in Figure 13, as we can precisely predict the misalignment in SAEC using the WGN as the reference signals [34,75]. We observe from Figure 15 that for a given ΔR, the coherence is generally inversely dependent on the signal frequency. In particular, we see that before the coherence reaches the first zero, the coherence reduction vs. frequency relationship is approximately linear. Furthermore, for a fixed frequency, e.g., at around 3 kHz, before the coherence first reaches zero, the coherence also decreases roughly linearly as ΔR increases. Thus (129) provides a way to measure the amount of decorrelation at each frequency point for a certain expansion ratio R. Conversely, it allows us to control R for a desired amount of decorrelation in terms of the coherence at certain frequency points, e.g., to minimize the distortion of a signal at low frequencies.

4.2.3 Misalignment vs. Resampling: without Far-End Room

We now show the link between the proposed resampling scheme in Figure 13 and the misalignment of SAEC when the reference signals of both channels are resampled

from a single WGN source. To fit the OLS FDAF structure, let L be the length of the adaptive filter and $2L$ be the length of a discrete STFT analysis frame. The analysis frame covers Q cycles of the sawtooth delay variation in Figure 13 such that the time-varying delay mismatch occurs more often within the frame. For simplicity, we assume Q to be a positive integer.

Let $u[n]$ be a zero-mean WGN with variance σ_u^2 and $\tilde{x}_p[n]$, $p = 1, 2$, be the resampled version of $u[n]$ by applying the resampling scheme in Figure 13. The following relationship can be established (see Appendix A for more details)

$$\tilde{X}_{p,m}[k] = \frac{1}{2L} \sum_{l=-L}^{L-1} U_m[l] \Phi_p(k, l), \quad p = 1, 2, \tag{130}$$

where $k = 0, \ldots, 2L-1$ and $U_m[l]$ is the l^{th} DFT coefficient of the windowed signal $u_m[n]$ in the m^{th} frame. Note that negative frequency indices, similarly used in (120), are used for a compact notation. $\Phi_p(k, l)$, $p = 1, 2$, is defined as

$$\Phi_1(k, l) \equiv \phi_N\big(\tfrac{1}{2Q}\big(\tfrac{l}{R} - k\big)\big)\phi_Q(l-k) + \phi_N\big(\tfrac{1}{2Q}(Rl - k)\big)\phi_Q(l-k)e^{j\frac{1}{Q}\pi[(1-\Delta R)l-k]} \tag{131}$$

$$\Phi_2(k, l) \equiv \phi_N\big(\tfrac{1}{2Q}(Rl - k)\big)\phi_Q(l-k) + \phi_N\big(\tfrac{1}{2Q}\big(\tfrac{l}{R} - k\big)\big)\phi_Q(l-k)e^{j\frac{1}{Q}\pi[(1+\frac{\Delta R}{R})l-k]}, \tag{132}$$

where $\phi_Q(x)$ is similarly defined as in (122). Using (130), (131), and (132) together with (125), the CSD, the PSDs, and the coherence can be written as

$$\hat{S}_{\tilde{x}_1\tilde{x}_2}[k] = \frac{\sigma_u^2}{2L} \sum_l \Phi_1(k, l)\Phi_2^*(k, l) \tag{133}$$

$$\hat{S}_{\tilde{x}_p\tilde{x}_p}[k] = \frac{\sigma_u^2}{2L} \sum_l \big|\Phi_p(k, l)\big|^2, \quad p = 1, 2 \tag{134}$$

$$\hat{C}_{\tilde{x}_1\tilde{x}_2}[k] = \frac{\big|\sum_l \Phi_1(k, l)\Phi_2^*(k, l)\big|^2}{\big(\sum_l \big|\Phi_1(k, l)\big|^2\big)\big(\sum_l \big|\Phi_2(k, l)\big|^2\big)}, \tag{135}$$

where $l = -L, \ldots, L-1$.

Using (134) and (135) with (100), we can obtain

$$\text{tr}\{\mathbf{S}^{-1}\} = \frac{2L}{\sigma_u^2} \sum_{k=0}^{2L-1} \Omega(R, k), \tag{136}$$

where (omitting the indices for simplicity)

$$\Omega \equiv \frac{\left(\sum_l \left|\Phi_1\right|^2\right) + \left(\sum_l \left|\Phi_2\right|^2\right)}{\left(\sum_l \left|\Phi_1\right|^2\right)\left(\sum_l \left|\Phi_2\right|^2\right) - \left|\sum_l \Phi_1 \Phi_2^*\right|^2}. \tag{137}$$

Using (136) and (137), the theoretical steady-state misalignment (91) can be analytically expressed as a function of R

$$\zeta(R) = \zeta_{\text{MSE}} + 10\log_{10}\left(L\sum_{k=0}^{2L-1}\Omega(R,k)\right), \tag{138}$$

where the excess mean squared error (MSE) or the lower bound of the misalignment is

$$\zeta_{\text{MSE}} = 10\log_{10}\left(\frac{1-\lambda}{\text{ENR}}\right) \tag{139}$$

and the echo-to-noise ratio (ENR) is $\|\mathbf{h}\|_2^2\sigma_u^2/\sigma_v^2$.

Figure 16 shows the misalignment vs. resampling ratio plot with $L = 1024$, $Q = 4$, $\lambda = (1 - 1/(6L))^L$, ENR = 25 and 35 dB, and the resampling ratio varying from 1 to 1.05. The straight lines represent the lower bounds of the misalignment when the two channels are uncorrelated, i.e., $\mathbf{C}_{12} = \mathbf{0}_{2L\times 2L}$. Note that the misalignment goes to infinity if $R = 1$, i.e., no resampling, since the spectral density matrix $\mathbf{S}$ becomes singular. As expected, the misalignment approaches the lower bound as the resampling ratio increases.

4.2.4 Misalignment vs. Resampling: with Far-End Room

We now show the effect of resampling on the reference signals that are obtained from convolving a single WGN source with the far-end room impulse response. Let $u[n]$ be the zero-mean WGN at the far-end room with variance σ_u^2. The noise is convolved with the far-end room impulse response $g_p[n]$, $p = 1, 2$. Assuming the length of the far-end room impulse response is $K \geq 2L$, the CSD of the reference signals after resampling is given by (see Appendix B for more details)

$$\hat{S}_{\tilde{x}_1\tilde{x}_2}[k] = \frac{\sigma_u^2}{8KL^2}\sum_{s=-K}^{K-1} G_1[s]\Psi_1(k,s)G_2^*[s]\Psi_2^*(k,s), \tag{140}$$

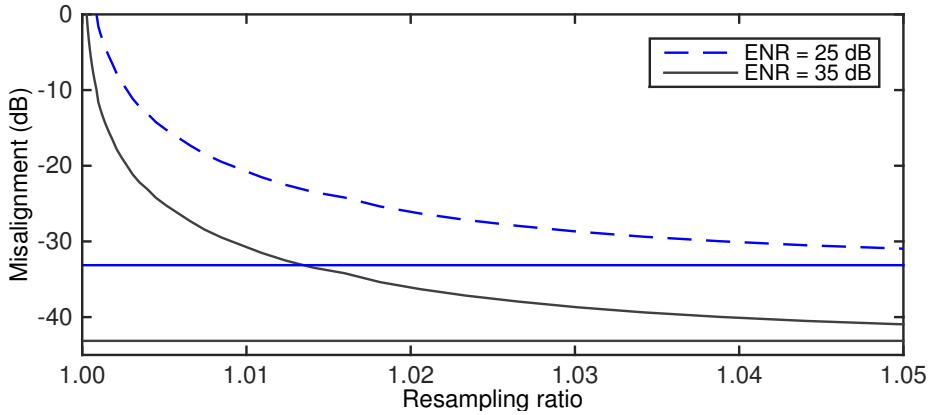


Figure 16: Misalignment vs. resampling ratio plot obtained from (138). The straight lines represent the lower bounds, i.e., ζ_{MSE}, of the misalignment when the two channels are uncorrelated.

where

$$G_p[s] = \sum_{n=0}^{K-1} g_p[n] e^{-j\frac{\pi}{K}sn}, \qquad p = 1, 2 \tag{141}$$

$$\Psi_p(k,s) \equiv \sum_{l=-L}^{L-1} \Phi_p(k,l)\phi_{2L}\big(\tfrac{L}{K}s - l\big), \qquad p = 1, 2 \tag{142}$$

with $k = 0, \ldots, 2L-1$ and $\phi_{2L}(x)$ defined similarly as in (122). Given (140), the PSDs and the coherence are given by

$$\hat{S}_{\tilde{x}_p\tilde{x}_p}[k] = \frac{\sigma_u^2}{8KL^2} \sum_s \big|G_p[s]\Psi_p(k,s)\big|^2, \quad p = 1, 2 \tag{143}$$

$$\hat{C}_{\tilde{x}_1\tilde{x}_2}[k] = \frac{\big|\sum_s G_1[s]\Psi_1(k,s)G_2^*[s]\Psi_2^*(k,s)\big|^2}{\big(\sum_s \big|G_1[s]\Psi_1(k,s)\big|^2\big)\big(\sum_s \big|G_2[s]\Psi_2(k,s)\big|^2\big)}, \tag{144}$$

where $s = -K, \ldots, K-1$.

Using (143) and (144) with (91) and (100), we can obtain the misalignment as a function of the resampling ratio

$$\zeta(R) = 10\log_{10}\left(\frac{(1-\lambda)\sigma_v^2}{\|\mathbf{h}\|_2^2\sigma_u^2} \cdot 4KL^2 \sum_{k=0}^{2L-1} \Theta(R,k)\right), \tag{145}$$

where (omitting the indices for simplicity)

$$\Theta \equiv \frac{\big(\sum_s \big|G_1\Psi_1\big|^2\big) + \big(\sum_s \big|G_2\Psi_2\big|^2\big)}{\big(\sum_s \big|G_1\Psi_1\big|^2\big)\big(\sum_s \big|G_2\Psi_2\big|^2\big) - \big|\sum_s G_1\Psi_1 G_2^*\Psi_2^*\big|^2}. \tag{146}$$

Figure 17 shows the misalignment vs. resampling ratio plot with $L = 1024$, $K = 2048$, $Q = 4$, $\lambda = (1 - 1/(6L))^L$, ENR = 25 and 35 dB, and the resampling ratio varying from 1 to 1.05. The measured far-end room impulse response with $T_{60} = 250$ ms was truncated to $K = 2048$ taps (128 ms). The ENR was measured using ENR $= \|\mathbf{h}\|_2^2\|\mathbf{g}\|_2^2\sigma_u^2/\sigma_v^2$ to take into account the far-end room impulse response. The lower bound is calculated with (145) by setting the coherence (144) to zero, i.e., by setting $\sum_s G_1\Psi_1 G_2^*\Psi_2^* = 0$ in (146). This is required since incorporating the far-end room impulse response in the model effectively raises the lower bound above ζ_{MSE}.

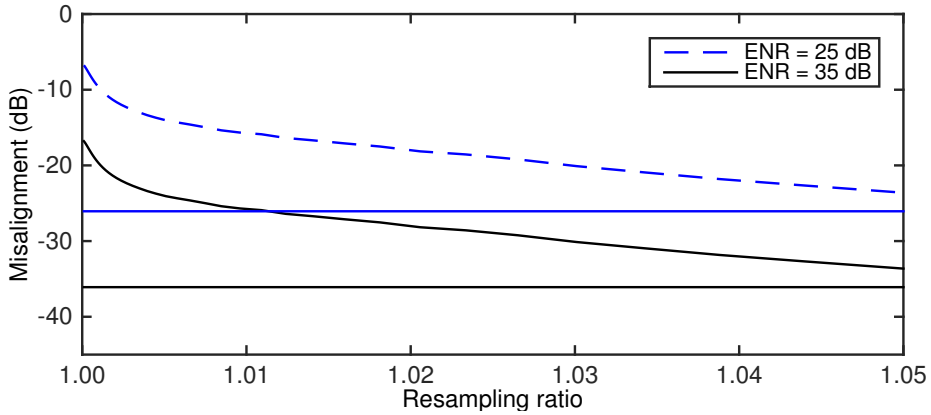


Figure 17: Misalignment vs. resampling ratio plot obtained from (145). The straight lines are the lower bounds of the misalignment when the coherence is zero.

Compared to Figure 16, Figure 17 shows that even without resampling, i.e., $R = 1$, the misalignment is lower than 0 dB. This is expected since the far-end room impulse response may slightly decorrelate the reference signals even though they are generated from the same source at the far-end room. On the other hand, the achievable misalignment at the same resampling ratio is less than (138) when we model the far-end room in (145). This is also expected since spectral nulls at the far-end room effectively lowers the PSDs at the spectral nulls. Due to the spectral nulls, the spectral density matrix $\mathbf{S}$ may be ill-conditioned, resulting in a higher misalignment according to (91) and (100). To achieve the same level of misalignment, a much larger resampling ratio

is required when we model the effect of the far-end room impulse response.

4.3 Algorithmic Design and Related Issues

In this section, we discuss the implementation details of the proposed resampling scheme. Proper attention must be paid when constructing the resampling scheme to achieve decorrelation without introducing an unnecessary time-domain or frequency-domain distortion. We present two frameworks for implementation, i.e., frequency-domain resampling (FDR) and time-domain resampling (TDR), and discuss the relative strengths and weaknesses of each framework. Finally, we discuss the trade-off in terms of resampling accuracy and processing delay for different frameworks.

4.3.1 Proper Resampling Scheme

Resampling a block of N samples introduces a total delay of $N(R-1)$ samples, where time expansion ($R > 1$) and time compression ($0 < R < 1$) introduce positive and negative instantaneous (sub-)sample delay, respectively. Since the discrete-time signal is resampled block-by-block without any overlap, there can potentially be a signal discontinuity between the blocks if we do not resample each block correctly.

Although a signal is usually resampled in one direction, i.e., forward in time, it may also be resampled in the backward direction by first time-reversing the signal block, applying the resampling procedure, and reversing the block back afterward. Different combinations of the resampling ratio (expansion or compression) and the resampling direction (forward or backward) give rise to four possibilities: forward expansion, forward compression, backward expansion, and backward compression. The change in the delay after resampling a signal block in four different situations are illustrated in Figure 18.

There are two constraints for smooth transition between the resampled blocks. First, there should be no sudden change in the delay across blocks. Second, the reference points of the adjacent blocks should be matched. In other words, we should

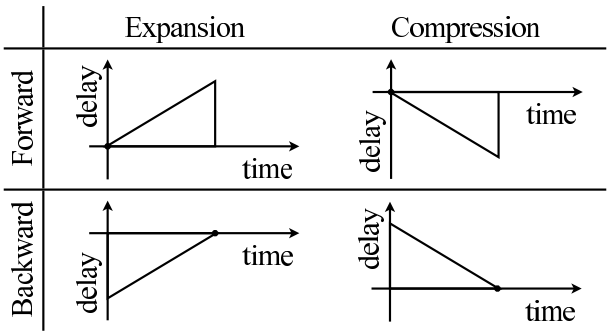
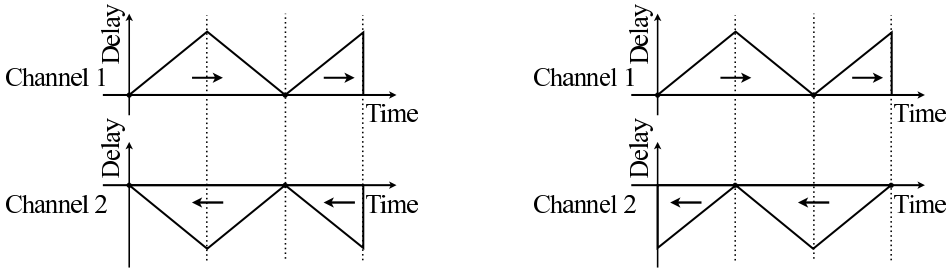


Figure 18: Signal delay after resampling. The black dots indicate the anchoring point from which the positive/negative delay starts to grow after resampling.

connect edges to edges and dots to dots in Figure 18 to ensure smooth transition of the resampling delay across blocks. Otherwise, a sudden change in the delay across block edges introduces a signal discontinuity, or decimation, which in turn causes the undesirable aliasing distortion [106].

Based on the delay smoothing rules, several possible resampling schemes can be designed, and Figure 19 shows two schemes that obey the delay continuity constraints. Although it may appear that the alternative scheme in Figure 19(b) can achieve inter-channel decorrelation, it actually fails to do so and thus should be avoided. The reason is that the expansion or the compression occurs in both channels at the same time, with the only difference being resampling in forward or backward direction. That is, expanding or compressing the channels simultaneously with the same resampling ratio R near unity results in a slight shifting of the entire blocks in the opposite time direction. Due to the constant amount of induced delay between two blocks, the CSD is unchanged and therefore no short-time decorrelation occurs. In other words, the *instantaneous delay difference* between channels is constant in such a case. The entire process becomes much like the input-sliding technique of [106] but with no aliasing distortion at all due to the delay smoothing, hence no decorrelation. For the proposed scheme, the delay difference between channel 1 and channel 2 continuously varies with time. This specifies another design rule, where for a given time, two

adjacent channels must not be expanded or compressed with the same R even if the direction of resampling is different.



(a) Proposed scheme. (b) Inappropriate alternative scheme.

Figure 19: Resampling schemes. The first scheme properly decorrelates the reference signals whereas the second one fails to do so.

4.3.2 Frequency-Domain Resampling

For continuous-time signals, we know from the time-scaling property of the CTFT that $\text{CTFT}\{x(t/R)\} = RX(R\omega)$. The goal of FDR is to interpolate across frequency rather than across time, with appropriate expansion or compression of the spectrum, to reduce the computation via the fast Fourier transform (FFT). Given a resampling ratio $0 < R < 2$, the FDR procedure for an N-point signal block of $x[n]$ is as follows [120]:

- Zero-extend the signal by a factor of $J = 2^{\alpha}$, $\alpha \geq 1$.
- Apply a JN-point FFT on the zero-extended signal.
- Linearly interpolate between $X[k]$ and $X[k+1]$

$$\tilde{X}[\tilde{k}] = R((1-\beta)X[k] + \beta X[k+1]) \tag{147}$$

  with the constraints $k \leq R\tilde{k} \leq k+1$ and $\beta = R\tilde{k} - k$ for each $\tilde{k}^{\text{th}}$ new sample to form $2N$ equally spaced samples.

- Apply a $2N$-point inverse fast Fourier transform (IFFT) on $\tilde{X}[\tilde{k}]$.

- Discard the samples at the end of the resampled signal $\tilde{x}[n]$ to retain the first RN resampled values.

Using the zero-extension factor $J \geq 2$ and taking the $2N$-point IFFT avoids the time domain aliasing after resampling with $R > 1$. Although it was shown that a zero-extension power $4 \leq \alpha \leq 6$ is sufficient in terms of the resampling accuracy for most applications [120], this translates to a rather large signal block size JN for $\alpha \geq 6$, where the computational complexity per block is dominated by the FFT and the complexity per sample is $\mathcal{O}(J \log(JN))$.

4.3.3 Time-Domain Resampling

According to (120), we can reformulate the TDR process by using the DFT matrix and matrix-vector multiplication. Let $\mathbf{x} = [x[0], \ldots, x[N-1]]^{\mathrm{T}}$, $\tilde{\mathbf{x}} = [\tilde{x}[0], \ldots, \tilde{x}[N-1]]^{\mathrm{T}}$, and $\mathbf{F}$ be the $N \times N$ DFT matrix, where $[\mathbf{F}]_{k+1,n+1} = e^{-j\omega_k n}$ is the element at the $(k+1)^{\text{th}}$ row and the $(n+1)^{\text{th}}$ column and $\omega_k = 2\pi k/N$. Let $\mathbf{M}$ be the MIDFT matrix with resampling ratio R. Each element in $\mathbf{M}$ is then given by

$$[\mathbf{M}]_{n+1,k+1} = \begin{cases} \frac{1}{N} e^{j\frac{\omega_k}{R}n}, & k \in [0, \frac{N}{2}] \text{ and } \frac{\omega_k}{R} < \pi \\ 0, & k \in [0, \frac{N}{2}] \text{ and } \frac{\omega_k}{R} > \pi \\ 0, & k \in [\frac{N}{2}+1, N-1] \text{ and } \frac{\omega_{N-k}}{R} > \pi \\ \frac{1}{N} e^{-j\frac{\omega_{N-k}}{R}n}, & k \in [\frac{N}{2}+1, N-1] \text{ and } \frac{\omega_{N-k}}{R} < \pi. \end{cases} \tag{148}$$

The low-pass filtering of $X[k]$ is directly embedded in the matrix $\mathbf{M}$ to account for possible aliasing when time compression occurs, i.e., when $R' = 1/R$ is used in stead of R. With the given DFT and MIDFT matrices, (120) becomes

$$\tilde{\mathbf{x}} = \mathbf{MFx} = \mathbf{Px}, \tag{149}$$

where $\mathbf{P} \equiv \mathbf{MF}$ is defined as the resampling matrix. Although $\mathbf{F}$ and $\mathbf{M}$ are in general complex, $\mathbf{P}$ is always real due to the conjugate symmetries in the rows of $\mathbf{M}$

and the columns of $\mathbf{F}$. For a fixed R, the resampling matrix can be computed once and stored in memory for future resampling tasks.

4.3.4 Block Processing

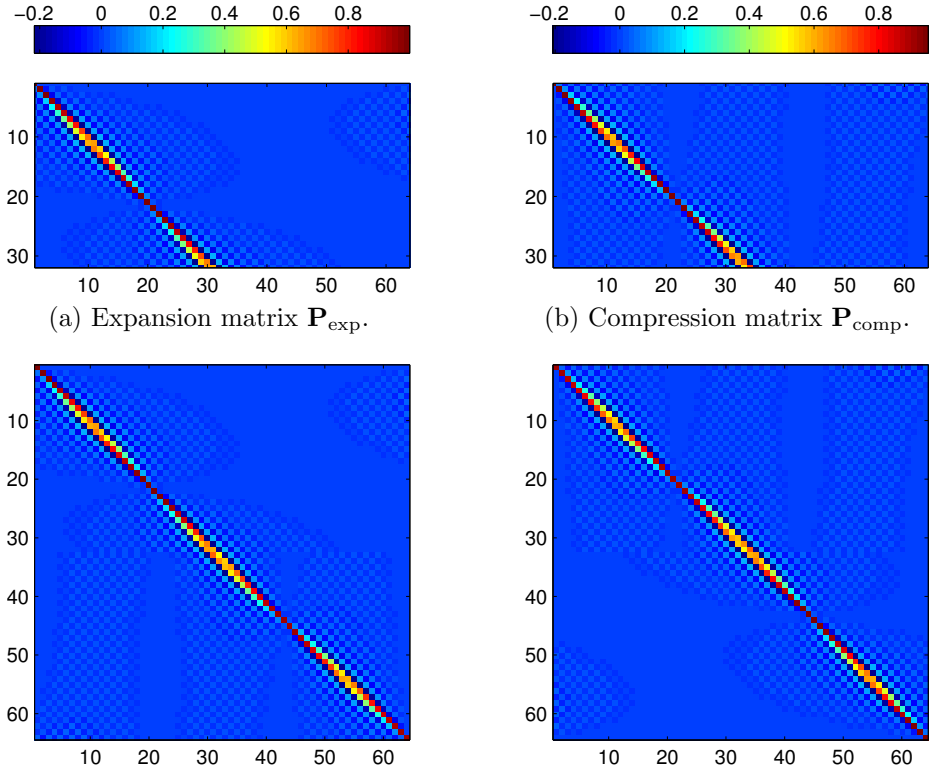
As shown in Figure 19(a), the resampling delay goes back to zero every two blocks, which naturally ensures continuity at those points. However, the continuity may not be maintained at other block boundaries where sudden changes, i.e., the non-zero inflection points in Figure 19(a), in the delay occur. To preserve the continuity, the resampling procedure must be constructed properly every two blocks. The following analysis focuses on TDR although similar principles to smooth out the delay should be applied to FDR as well.

Let $\mathbf{x}_{2N} = [x[0], x[1], \ldots, x[2N-1]]^{\mathrm{T}}$ represent a block of signal, where $2N$ is the block size. We now construct a $2N \times 2N$ matrix $\mathbf{M}$ and apply an FFT of size $2N$ on each row of the matrix to obtain $\mathbf{P}$, where we form one resampling matrix for expansion and another for compression. The rows of the two matrices represent a set of interpolation filters for either signal expansion or compression.

Figure 20 shows the surface plots corresponding to the resampling matrices with $N = 32$ and $R = 1.0512$ for two channels. The upper halves of the expansion and the compression matrices are shown in Figures 20(a) and 20(b), denoted by $\mathbf{P}_{\mathrm{exp}}$ and $\mathbf{P}_{\mathrm{comp}}$, respectively. They are the relevant portions for the construction of the actual resampling matrices for the two channels as shown in Figures 20(c) and 20(d), denoted by $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively, and obtained by

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{P}_{\mathrm{exp}} \\ (\mathbf{P}_{\mathrm{comp}}^{\dagger})_{2N}^{\hookrightarrow} \end{bmatrix} \text{ and } \mathbf{P}_2 = \begin{bmatrix} \mathbf{P}_{\mathrm{comp}} \\ (\mathbf{P}_{\mathrm{exp}}^{\dagger})_{2N}^{\hookrightarrow} \end{bmatrix}, \tag{150}$$

where $\{\cdot\}^{\dagger}$ is a flipping operator that flips a matrix both horizontally and vertically and $\{\cdot\}_{2N}^{\hookrightarrow}$ denotes circular shifting the columns to the right by $2N$. The reason for the circular shifting will become obvious later, which for now does not change



(a) Expansion matrix $\mathbf{P}_{\text{exp}}$.

(b) Compression matrix $\mathbf{P}_{\text{comp}}$.

(c) Resampling matrix $\mathbf{P}_1$.

(d) Resampling matrix $\mathbf{P}_2$.

Figure 20: Resampling matrices with $N = 32$ and $R = 1.0512$ for channels 1 and 2. Note that a larger N and a smaller R are typically used. These parameters are chosen here for illustration purpose.

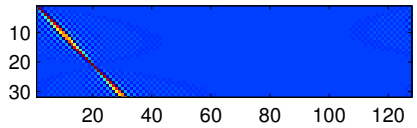
the resulting resampling matrices. By combining each half of the expansion and the compression matrices, the continuity is ensured after resampling as indicated by the smooth transition of the coefficients for the interpolation filters (i.e., horizontal cross-sections) of $\mathbf{P}_1$ and $\mathbf{P}_2$ in Figures 20(c) and 20(d). The corresponding computational complexity per sample is $\mathcal{O}(N)$ with an algorithmic delay of $2N$ samples.

4.3.4.1 Block Mirroring

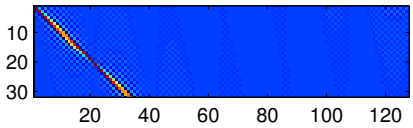
Due to the nature of block processing by the proposed resampling scheme and the spectral leakage, the resampling error at the block edges will inevitably be higher than towards the center of the block. More specifically, the DFT assumes the input signal to be periodic, but periodically repeated blocks of any signal are not guaranteed to be

continuous. We may then extend the signal block by mirroring it to make the block edges circularly continuous and to improve the resampling accuracy.

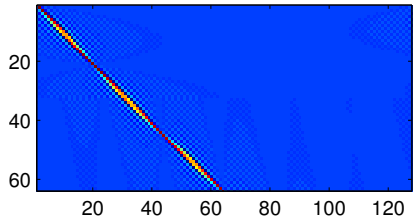
Let $\mathbf{x}_{2N}^{\mathrm{M}} = [\mathbf{x}_{2N}^{\mathrm{T}}, (\mathbf{x}_{2N}^{\mathrm{T}})^{\dagger}]^{\mathrm{T}}$ be formed by concatenating $\mathbf{x}_{2N}$ with its mirrored version. Since the block size is $4N$, the resampling matrix is constructed through the FFT of size $4N$. However, we are only interested in the first $2N$ samples after resampling $\mathbf{x}_{2N}^{\mathrm{M}}$. Therefore, it suffices to take the upper quarters, i.e., the $N \times 4N$ sub-matrices, of the mirrored resampling matrices corresponding to expansion and compression. This is illustrated in Figure 21, where $\mathbf{P}_{\mathrm{exp}}^{\mathrm{M}}$ and $\mathbf{P}_{\mathrm{comp}}^{\mathrm{M}}$, shown in Figures 21(a) and 21(b), respectively, are used to construct via (150) the resampling matrices $\mathbf{P}_{1}^{\mathrm{M}}$ and $\mathbf{P}_{2}^{\mathrm{M}}$, shown in Figures 21(c) and 21(d), respectively. We note here that the circular shifting in (150) is indeed required in this case to ensure that the coefficients for the interpolation filters are continuous. The computational complexity per sample is $\mathcal{O}(N)$ with an algorithmic delay of $2N$ samples.



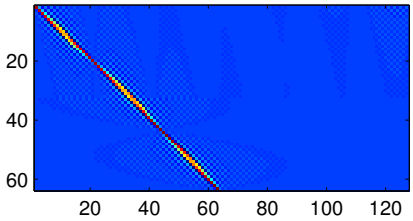
(a) Expansion matrix $\mathbf{P}_{\mathrm{exp}}^{\mathrm{M}}$.



(b) Compression matrix $\mathbf{P}_{\mathrm{comp}}^{\mathrm{M}}$.



(c) Resampling matrix $\mathbf{P}_{1}^{\mathrm{M}}$.



(d) Resampling matrix $\mathbf{P}_{2}^{\mathrm{M}}$.

Figure 21: Resampling matrices for the mirrored signal.

4.3.4.2 Center Filtering

The mirroring may be less than ideal since we do not have access to the future signal values after the latest block. By utilizing a look-ahead strategy, we can further reduce the block edge distortion and the processing delay by center filtering. Observing the

structure of the resampling matrices in Figure 20, we can immediately see that only the center rows of the resampling matrices achieve the center filtering of the input signal block $\mathbf{x}_{2N}$. All the other rows of the resampling matrices correspond to filtering a periodic signal $\mathbf{x}_{2N}$ that is not equal to the actual underlying input signal as the periodicity is assumed by the DFT. Therefore, instead of processing the input signal block-wise, we may, in the same manner as resampling by interpolation, resample continuously in time by applying the appropriate interpolation filter to generate each sample. This is achieved by circularly shifting the rows of the resampling matrices such that all of the coefficients for the interpolation filters are appropriately centered at the middle columns of the matrices.

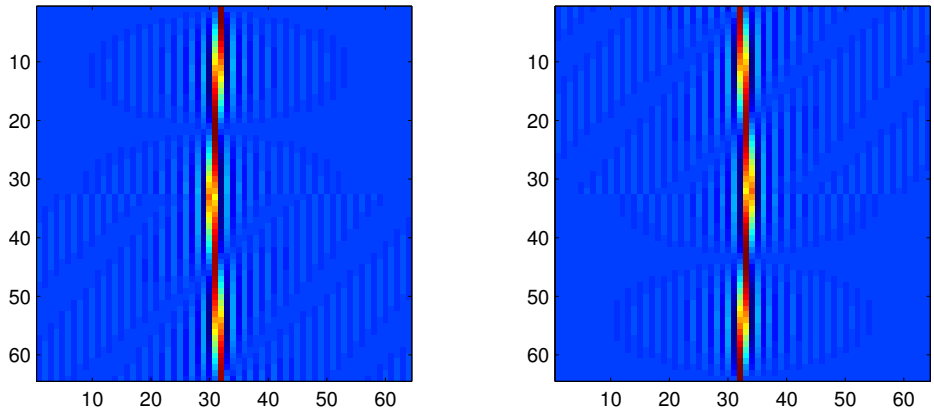
Figure 22 shows the resampling matrices, denoted by $\mathbf{P}_1^{\text{CIRC}}$ and $\mathbf{P}_2^{\text{CIRC}}$ for the two channels, after circular shifting. The input signal block to be used with the circularly shifted resampling matrices is

$$\mathbf{x}_{2N}^{\text{C}} = [x[-N], \ldots, x[-1], x[0], \ldots, x[N-1]]^{\text{T}}, \tag{151}$$

with $\{x[n]; n < 0\}$ being the past samples, $x[0]$ being the current sample, and $\{x[n]; n > 0\}$ being the future samples. For subsequent rows of the resampling matrix, the input signal block is shifted to the right one sample at a time to ensure center filtering. The computational complexity per sample is still $\mathcal{O}(N)$ but the algorithmic delay is lowered to N samples.

4.3.5 Comparison of Frequency- and Time-Domain Resampling

The computational complexity for different resampling methods is summarized in Table 14. The resampling accuracy of FDR and TDR is compared using sinusoids. Given a sine wave $x[n] = \sin(2\pi \frac{f}{f_s} n)$, where f is the sinusoidal frequency and $f_s =$ 16 kHz is the sampling frequency, the resampled signal $\tilde{x}[n]$ is compared to the ground truth signal $\bar{x}[n] = \sin(2\pi \frac{f}{Rf_s} n)$ with a resampling ratio $R = 1.001$. The sinusoidal frequency is swept from 20 Hz to 8 kHz with an increment of 10 Hz. The block size is



(a) Resampling matrix $\mathbf{P}_1^{\text{CIRC}}$. (b) Resampling matrix $\mathbf{P}_2^{\text{CIRC}}$.

Figure 22: Resampling matrices after circular shifting.

set at $N = 512$. To evaluate the resampling accuracy, the signal-to-error ratio (SER), defined below, is used.

$$\text{SER} = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} x^2[n]}{\sum_{n=0}^{N-1} (\tilde{x}[n] - \bar{x}[n])^2} \right). \tag{152}$$

Table 14: Complexity per sample and algorithmic delay comparison.

Method	Complexity	Delay
FDR, $J = 2^\alpha$	$\mathcal{O}(J \log(JN))$	$2N$
Mirroring	$\mathcal{O}(N)$	$2N$
Center Filtering	$\mathcal{O}(N)$	N

Figure 23 shows the SER obtained from different resampling methods. For FDR to have high enough accuracy, α has to be as high as 6. Although not shown, the resampling accuracy for FDR cannot be improved much further when α is increased beyond 6. TDR, on the other hand, gives higher accuracy than FDR at $\alpha = 6$ and is used for the experimental evaluation.

The mirroring method and the center filtering method achieve the highest resampling accuracy in general. We note, however, that the resampled signal is a simple sinusoid, where the block edge distortion will not be high since the mirrored signal

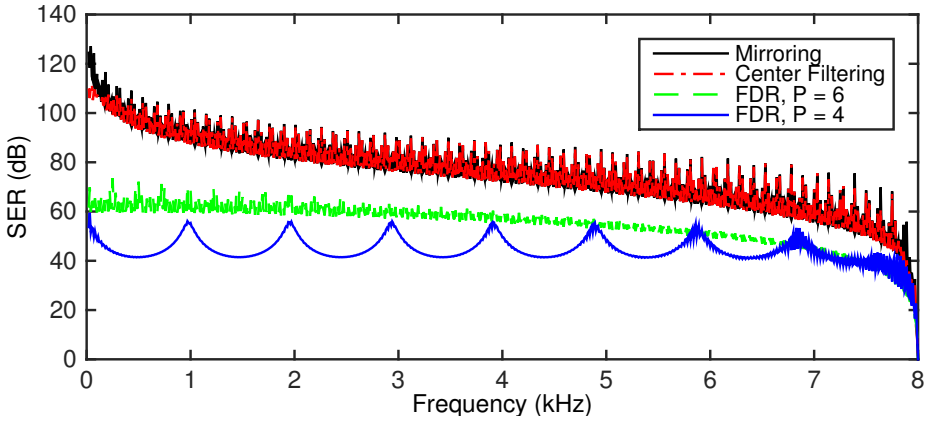


Figure 23: SER plot for different resampling methods.

still has the same frequency. For actual signals such as speech, we expect that the mirroring method will result in lower resampling accuracy than the centering filtering method.

Although TDR has comparable complexity and delay to FDR, the construction of the resampling matrix $\mathbf{P}$ for TDR, however, can be computationally expensive since it requires a complexity of $\mathcal{O}(N^3)$ for the multiplication of two $N \times N$ matrices in (149). On the other hand, we note that the matrix product $\mathbf{MF}$ is equivalent to applying the DFT row-wise on the matrix $\mathbf{M}$. Thus instead of directly computing the matrix product, we can apply the FFT on each row of $\mathbf{M}$ to obtain $\mathbf{P}$. This reduces the complexity in practice to $\mathcal{O}(N^2 \log N)$. For example, with a sampling rate of $f_s = 16000$ kHz and a typical block size of $N = 512$, the computational saving can be roughly 57 fold. For a fixed R the computation may be required to be done only once, but the savings still translate to significantly faster initialization of the resampling procedure.

In general TDR outperforms FDR in terms of the resampling accuracy at the cost of high complexity $\mathcal{O}(N^2 \log N)$ for the resampling matrix initialization. Depending on applications, this high complexity may prevent us from changing the resampling ratio on the fly, where FDR becomes the preferred choice for this task. On the other hand, we may precompute a number of resampling matrices and store the results, so

that we can still retain the resampling accuracy of TDR while having the flexibility of changing the resampling ratio on the fly at the same time.

4.3.6 Sub-Band Resampling

For the perceptual quality and the actual cancellation performance reasons [118,120], we may want to modify the signal only in certain sub-bands [124]. To that end, Figure 15 suggests that to achieve the same overall reduction in the coherence, the resampling ratio R may be adjusted separately over each sub-band in the frequency domain. This can be done to make sure that the speech distortion will be minimized by the resampling process. In addition, a sudden change in R between sub-bands, e.g., $R = 1$ in the low sub-band and $R > 1$ in the high sub-band, may introduce unwanted frequency-domain distortions. We experimentally verified that the distortion created by such a discontinuity in R has the characteristics of a musical noise. Therefore, we propose to vary the resampling ratio per frequency bin as smoothly across the bins as possible. This procedure involves making R a continuous function of frequency, i.e., $R[k]$, and using the desired $R[k]$ for resampling.

4.4 Experimental Evaluation

4.4.1 Application to Stereophonic Acoustic Echo Cancellation

A two-channel FDAF [5,34,75] is used to verify the accuracy of the theoretical steady-state misalignment in (145). Table 15 summarizes the FDAF algorithm. The regularization parameter plays an important role in adaptive algorithms [8]. Without proper regularization, an adaptive algorithm may not behave properly. Assuming a fixed regularization term δ, a constant term $\delta/(1-\lambda)$ should be added to the PSDs in (134) and (143) to reflect the regularization procedure in Table 15.

The room impulse responses were recorded at a 16 kHz sampling rate with a length of 4096. The near-end room impulse response was truncated to 1024 samples to neglect the effect on misalignment due to under-modeling, and the far-end room

Table 15: The two-channel FDAF [5].

Definitions

$$
\begin{aligned}
[\mathbf{F}]_{k+1,n+1} &= e^{-j\frac{\pi}{L}kn}, \quad k,n = 0,\ldots,2L-1 \\
\mathbf{G}^{01} &= \mathbf{F}\begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\end{bmatrix}\mathbf{F}^{-1} \\
\hat{\underline{\mathbf{h}}}_p &= \mathbf{F}[\hat{\mathbf{h}}_p^{\mathrm{T}}, \mathbf{0}_{L\times 1}^{\mathrm{T}}]^{\mathrm{T}}, \quad p = 1,2 \\
\mu' &= \mu(1-\lambda),\ 0 < \mu \leq 1,\ 0 \ll \lambda < 1
\end{aligned}
$$

Spectral estimation

$$
\begin{aligned}
\mathbf{X}_p[m] &= \mathrm{diag}\{\mathbf{F}[x_p[(m-1)L],\ldots,x_p[(m+1)L-1]]^{\mathrm{T}}\}, \quad p = 1,2 \\
\hat{\mathbf{S}}_{ij}[m] &= \lambda\hat{\mathbf{S}}_{ij}[m-1] + (1-\lambda)\mathbf{X}_i[m]\mathbf{X}_j^*[m], \quad i,j = 1,2 \\
\hat{\mathbf{S}}_{pp}[m] &= \hat{\mathbf{S}}_{pp}[m] + \delta\mathbf{I}_{2L\times 2L}, \quad p = 1,2 \\
\hat{\mathbf{C}}_{12}[m] &= (\hat{\mathbf{S}}_{11}[m]\hat{\mathbf{S}}_{22}[m])^{-1}\hat{\mathbf{S}}_{12}[m]\hat{\mathbf{S}}_{21}[m] \\
\hat{\mathbf{S}}_p[m] &= \hat{\mathbf{S}}_{pp}[m](\mathbf{I}_{2L\times 2L} - \hat{\mathbf{C}}_{12}[m]), \quad p = 1,2 \\
\mathbf{K}_1[m] &= \hat{\mathbf{S}}_1^{-1}[m](\mathbf{X}_1[m] - \hat{\mathbf{S}}_{12}[m]\hat{\mathbf{S}}_{22}^{-1}[m]\mathbf{X}_2[m]) \\
\mathbf{K}_2[m] &= \hat{\mathbf{S}}_2^{-1}[m](\mathbf{X}_2[m] - \hat{\mathbf{S}}_{21}[m]\hat{\mathbf{S}}_{11}^{-1}[m]\mathbf{X}_1[m])
\end{aligned}
$$

Filter adaptation

$$
\begin{aligned}
\underline{\mathbf{y}}[m] &= \mathbf{F}[\mathbf{0}_{1\times L}, y[mL],\ldots,y[(m+1)L-1]]^{\mathrm{T}} \\
\underline{\mathbf{e}}[m] &= \underline{\mathbf{y}}[m] - \mathbf{G}^{01}(\mathbf{X}_1[m]\hat{\mathbf{h}}_1[m-1] + \mathbf{X}_2[m]\hat{\mathbf{h}}_2[m-1]) \\
\hat{\underline{\mathbf{h}}}_p[m] &= \hat{\underline{\mathbf{h}}}_p[m-1] + 2\mu'\mathbf{G}^{10}\mathbf{K}_p^*[m]\underline{\mathbf{e}}[m], \quad p = 1,2
\end{aligned}
$$

impulse response was truncated to $K = 2048$. The length of the adaptive filter was $L = 1024$ with $\lambda = (1 - 1/(3L))^L$, $\mu = 1$, and $\delta = \frac{L(1+\sqrt{1+\mathrm{ENR}})}{\mathrm{ENR}}(\sigma_{x_1}^2 + \sigma_{x_2}^2)$ [8]. An uncorrelated WGN at ENR = 30 dB was added to the near-end microphone signal. The proposed resampling scheme in Figure 13 was applied to the two reference signals, which came from either a single WGN or a speech signal at the far-end room. The resampling block size was $N = 256$ with $Q = 4$ such that the reference signals frame of length $2L$ covered 4 cycles of the sawtooth wave.

The speech signal was obtained by concatenating roughly 200 randomly chosen utterances from the TIMIT [38] testing database and was modeled using the linear

predictive coding (LPC) analysis. The order of the LPC analysis was 20, and the analysis frame size was 320 samples with a 50% overlap. A Hamming window was applied to each LPC analysis frame before calculating the autocorrelation function. The averaged gain and the averaged spectrum of the speech signal were calculated from averaging the autocorrelation functions of all analysis frames and performing the LPC analysis on the averaged autocorrelation function. The theoretical steady-state misalignment was still calculated from (145) with a modified far-end room impulse response $\tilde{G}_p[k]$, $p = 1, 2$, to take into account the effect of the speech source. The modified far-end room impulse response was obtained from multiplying the original far-end room impulse response $G_p[k]$, $p = 1, 2$, with the averaged speech spectrum and the averaged gain in each frequency bin. The variance σ_u^2 in (140) and (143) was the variance of the excitation signal obtained from the LPC analysis of the speech signal.

Figure 24 shows the theoretical steady-state misalignment calculated from (145) and the actual measured misalignment using a single WGN source at the far-end room with the proposed resampling scheme applied to the reference signals. Note that a long period of time is required to achieve the theoretical lower bound with a relatively small ΔR, i.e., $\Delta R = 0.002$. The long convergence time for a small ΔR is expected since only a fixed far-end room impulse response is used for the simulation such that the spectral nulls in the far-end room impulse response effectively limits the convergence speed when only a small amount of decorrelation is applied. Furthermore, since we know the true ENR, the ideal regularization term sets a much smaller lower bound than without the ideal regularization such that the convergence time is long when only a small amount of decorrelation is applied. Nonetheless, the theoretical steady-state misalignment (145) accurately predicts the actual measured misalignment if we allow sufficient time for the adaptive filter to converge.

Figure 25 shows the misalignment measured with a single speech source at the

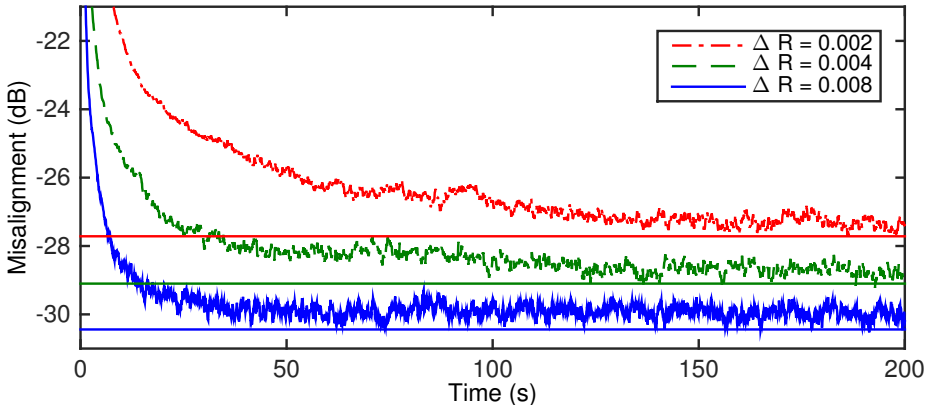


Figure 24: Misalignment for a WGN at the far-end room with the proposed resampling scheme. The straight lines are calculated from (145).

far-end room. Only two resampling ratios are shown to better illustrate the two misalignment curves since the theoretical steady-state misalignments are quite close together when using a speech source at the far-end room. We observe that the theoretical steady-state misalignment (145) with the modified far-end room impulse response $\tilde{G}_p[k]$, $p = 1, 2$, accurately predicts the lower bound of the actual misalignment. Note that for the same resampling ratio, the lower bound of the misalignment is larger for the speech source than the WGN. This is expected since the power spectrum of speech has a spectral roll-off of 6-10 dB/octave so that the high frequencies are weakly excited compared to the WGN. Even though the steady-state misalignment differs by at most 2 dB in Figures 24 and 25, the convergence rate is much faster for a larger ΔR. The convergence rate, rather than the lower bound of the misalignment, is an essential design criterion for choosing a proper resampling ratio.

4.4.2 Decorrelation by Sub-Band Resampling

In this section, we discuss how we properly choose the resampling ratios in different frequency bins to achieve fast convergence and a high speech quality after SBR. Since the energy of speech is usually concentrated below 4 kHz, we can divide the speech spectrum into several sub-bands, where we apply less decorrelation in the low frequency sub-bands to preserve the speech quality and more decorrelation in the high

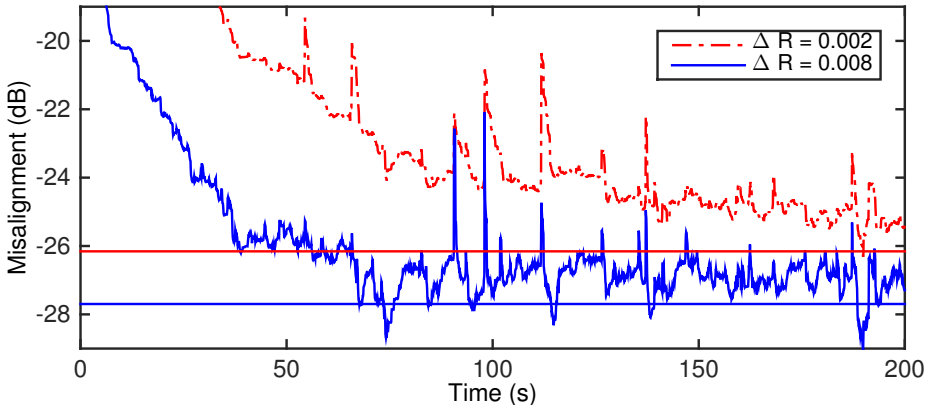


Figure 25: Misalignment for a speech signal at the far-end room with the proposed resampling scheme. The straight lines are the theoretical steady-state misalignment (145) with a modified far-end room impulse response.

frequency sub-bands to achieve a better convergence rate. This frequency-selective decorrelation strategy is consistent with the human sound localization capability in that the interaural level difference (ILD), rather than the interaural time difference (ITD), is used for localizing high frequency sounds and vice versa. ILD is not as susceptible to resampling as ITD.

Figure 26 shows the proposed SBR curve $\Delta R(f)$ for a sampling rate $f_s = 16$ kHz. Here we divide the frequency into three sub-bands, and only four resampling ratios, i.e., R_1 to R_4, need to be determined. We note that other possible SBR curves can be used, and this proposed SBR curve is chosen to reduce the number of controlling parameters. The resampling ratios are varied linearly within each sub-band. Note that ΔR_1 should be small so that less signal modification is applied below $f_s/4$ while ΔR_2 and ΔR_3 should be large to apply more resampling above $f_s/4$. ΔR_4 should be smaller than ΔR_3 since a relatively smaller ΔR_4 is sufficient for heavy decorrelation near the Nyquist frequency.

To determine the four resampling ratios in Figure 26, we first choose ΔR_1 to be relatively small, i.e., $\Delta R_1 = 0.001$, and determine ΔR_2 for a certain desirable coherence reduction in the high frequency sub-bands. In this experiment, we choose to fix the coherence to be below 0.8, 0.6, 0.4, 0.2, or 0.1 for frequencies above $f_s/4$,

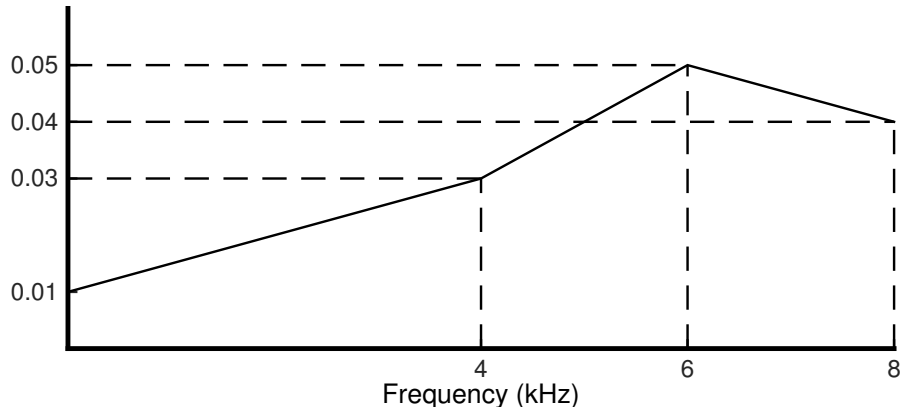


Figure 26: Proposed SBR curve for decorrelation. Note that the ΔR values are for illustration purpose and may be different depending on the design criteria.

and calculate the minimum required ΔR_2 using (135). ΔR_3 and ΔR_4 are chosen to be relatively large, i.e., $\Delta R_3 = 0.005$ and $\Delta R_4 = 0.004$, such that the coherence drops to zero from $3f_s/8$ to $f_s/2$. Figure 27 shows the required ΔR_2 to achieve the desired coherence in the high frequency sub-bands. We note that while (144) accurately calculates the coherence when the far-end room impulse response is involved, (135) is sufficient to determine the averaged behavior of the coherence. This is verified in Figure 28 by comparing (135) and (144) using the same SBR curve.

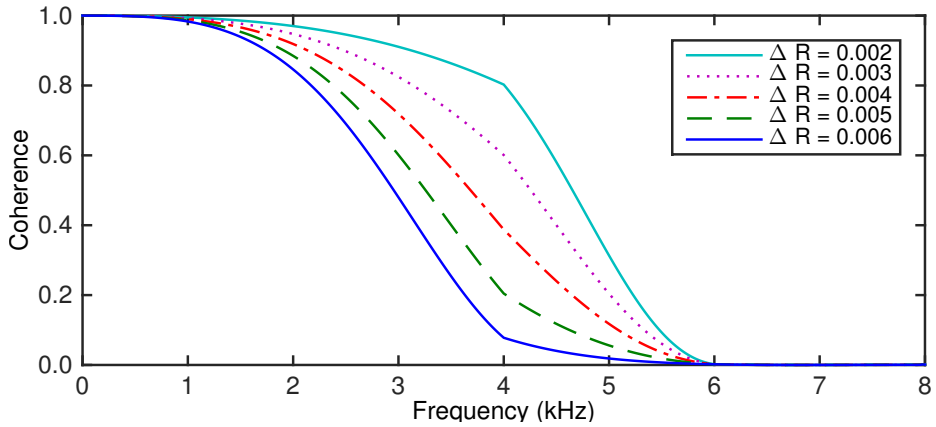


Figure 27: Coherence plot using (135) and the proposed SBR curve with $\Delta R_1 = 0.001$, $\Delta R_3 = 0.005$, and $\Delta R_4 = 0.004$, while varying ΔR_2 to achieve the desired coherence in the high frequency sub-bands.

Table 16 shows the speech quality measured by the Performance Evaluation of Speech Quality (PESQ) [71] with various values of ΔR_2, where PESQ^{NB} and PESQ^{WB}

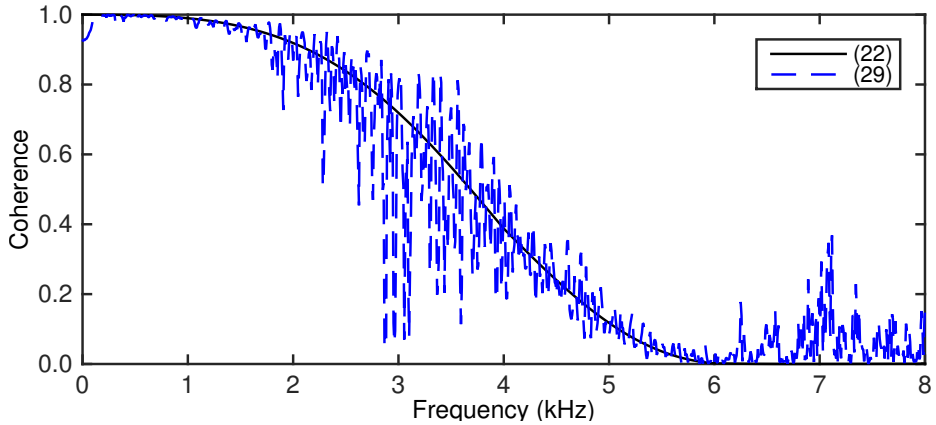


Figure 28: Comparison of (135) and (144) using the proposed SBR curve with $\Delta R_1 = 0.001$, $\Delta R_2 = 0.004$, and $\Delta R_3 = 0.005$, and $\Delta R_4 = 0.004$.

stand for the narrowband mode for handset listening and the wideband mode for headphone listening, respectively. The same utterances and the far-end room impulse response from Section 4.4.1 were used for the PESQ evaluation. Note that both PESQ^{NB} and PESQ^{WB} drop consistently as ΔR_2 increases. While varying ΔR_2 from 0.002 to 0.005 slightly degrades the speech quality, the degradation becomes more severe at $\Delta R_2 = 0.006$.

Table 16: Speech quality after applying the proposed SBR curve with $\Delta R_1 = 0.001$, $\Delta R_3 = 0.005$, and $\Delta R_4 = 0.004$.

ΔR_2	0.002	0.003	0.004	0.005	0.006
PESQ^{NB}	**4.47**	4.46	4.45	4.41	4.27
PESQ^{WB}	4.44	**4.45**	4.42	4.38	4.26

Figure 29 shows the misalignment plot with various values of ΔR_2. The measured misalignment curves are above the lower bounds (not shown) since the plot is zoomed in to better illustrate the convergence behavior. We observe that the convergence rate is dramatically improved for a larger ΔR_2, with a convergence performance gain of up to 5 dB. Note that while increasing ΔR_2 from 0.002 to 0.004 drastically improves the convergence rate, further increase of ΔR_2 shows marginal improvement while heavily degrading the PESQ. Due to these observations, we fix $\Delta R_2 = 0.004$ to achieve a sufficient convergence rate without sacrificing the speech quality too much.

Recall from Figure 27 that these values correspond to a coherence that is below 0.4 for frequencies above $f_s/4$ and gradually decreases to 0 at $3f_s/8$. Therefore, by properly choosing the resampling ratios in the high frequency sub-bands, we can heavily decorrelate the signal while introducing perceptually negligible distortion, as measured by the PESQ. Once the resampling ratios in the high frequency sub-bands are determined, we can vary ΔR_1 to achieve even more decorrelation in the low frequency sub-band and look at the convergence behavior and the processed speech quality.

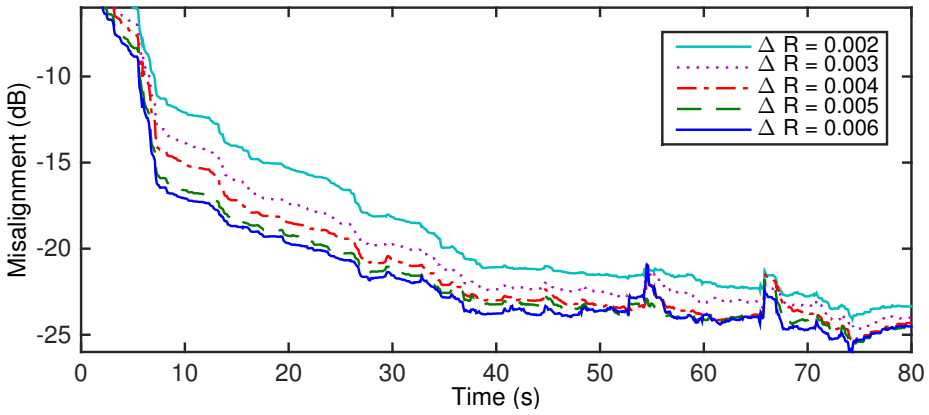


Figure 29: Misalignment for the proposed SBR curve with $\Delta R_1 = 0.001$, $\Delta R_3 = 0.005$, $\Delta R_4 = 0.004$, and various values of ΔR_2.

Table 17 and Figure 30 show the PESQ and the measured misalignment, respectively, with various values of ΔR_1. We observe from Table 17 that PESQ^{NB} stays high while PESQ^{WB} degrades slightly as ΔR_1 increases. Depending on the quality of the playback equipment, we may choose a larger ΔR_1 without harming the signal quality according to PESQ^{NB}. Here we base our choice on PESQ^{WB} assuming that high quality playback equipment is used. We observe that while the lower bounds are essentially the same for different resampling values of ΔR_1, a larger ΔR_1 results in much faster convergence. We further note from Figure 30 that while increasing ΔR_1 from 0.001 to 0.003 gives a nice boost to the convergence rate, increasing ΔR_1 beyond 0.003 provides marginal improvement while degrading PESQ^{WB}. From these

observations we choose $\Delta R_1 = 0.003$ for the improved convergence performance with slightly degraded PESQ^{WB}.

Table 17: Speech quality after applying the proposed SBR curve with $\Delta R_2 = 0.004$, $\Delta R_3 = 0.005$, and $\Delta R_4 = 0.004$.

ΔR_1	0.001	0.002	0.003	0.004	0.005
PESQ^{NB}	4.45	4.45	4.45	4.45	4.45
PESQ^{WB}	**4.42**	4.41	4.40	4.39	4.39

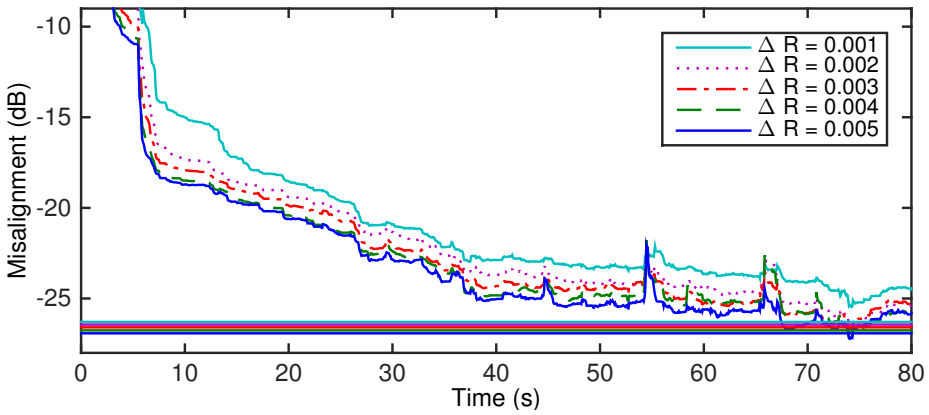


Figure 30: Misalignment for the proposed SBR curve with $\Delta R_2 = 0.004$, $\Delta R_3 = 0.005$, $\Delta R_4 = 0.004$, and various values of ΔR_1. The straight lines at the bottom represent the theoretical steady-state misalignment (145).

4.4.3 Comparison with Other Decorrelation Methods

To compare the SAEC performance with the proposed SBR curve against other commonly used decorrelation procedures, the following methods were tested:

- Uncorrelated additive white Gaussian noise (AWGN) at 30 dB signal-to-noise ratio (SNR).

- Nonlinear processor (NLP) [7, 34, 89], given by

$$x_1'[n] = x_1[n] + \frac{\gamma}{2}(x_1[n] + |x_1[n]|) \tag{153}$$

$$x_2'[n] = x_2[n] + \frac{\gamma}{2}(x_2[n] - |x_2[n]|), \tag{154}$$

  where $\gamma = 0.5$.

- Phase modulation (PMod) described in [55, Figure 2], where the window length was 48 at $f_s = 16$ kHz, the modulation frequency was 0.75 Hz, and the modulation amplitude was chosen according to [55, Figure 3].

- Proposed SBR curve with $\Delta R_1 = 0.003$, $\Delta R_3 = 0.005$, and $\Delta R_2 = \Delta R_4 = 0.004$.

Table 18 summarizes the quality measures using the segmental signal-to-noise ratio (SSNR), the log-spectral distortion (LSD), and the PESQ. $\text{PESQ}^{\text{LR-NB}}$ and $\text{PESQ}^{\text{LR-WB}}$ correspond to the evaluation obtained after averaging the measures taken individually from the left and the right channels. Although the SSNR and the LSD may not directly relate to the perceptual quality, the SSNR and the LSD measure the deviation of the processed signal from the original signal in the time domain and the frequency domain, respectively. We note that even though the SSNR and the LSD of AWGN are the best, the distortion introduced by AWGN at 30 dB SNR is quite audible as indicated by the low PESQ. Therefore, the SSNR or the LSD alone is not indicative of the perceptual quality after decorrelation, and the PESQ is a much more suitable measure for the perceptual quality.

Table 18: Processed speech quality comparison.

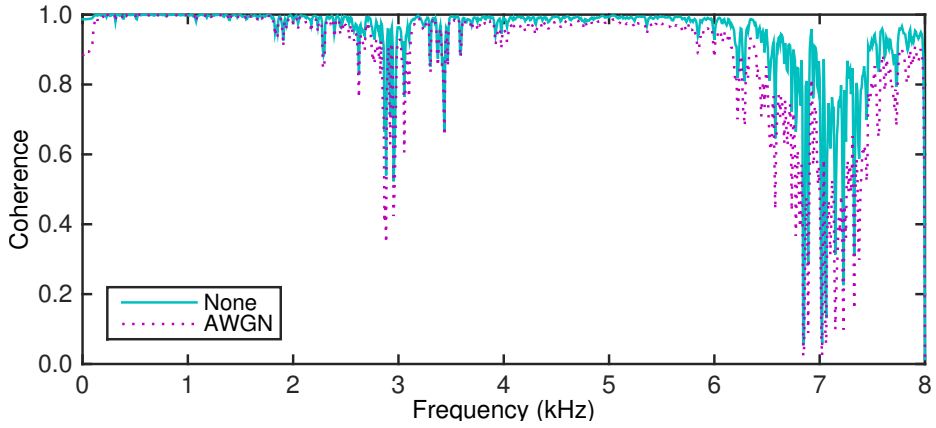
method	AWGN	NLP	PMod	SBR
SSNR (dB)	**18.31**	9.05	11.87	11.91
LSD (dB)	**0.13**	2.34	0.17	0.51
$\text{PESQ}^{\text{LR-NB}}$	3.90	3.82	**4.53**	4.52
$\text{PESQ}^{\text{LR-WB}}$	3.22	3.33	4.58	**4.59**
PESQ^{NB}	3.38	4.05	4.19	**4.45**
PESQ^{WB}	2.90	3.62	4.03	**4.40**

NLP has the highest LSD due to the heavy distortion introduced by the half-wave rectifier, and the frequency-domain distortion of NLP is clearly audible. The somewhat higher LSD of SBR is expected since the frequency contents (both the magnitude and the phase) are slightly shifted by the resampling process. The LSD

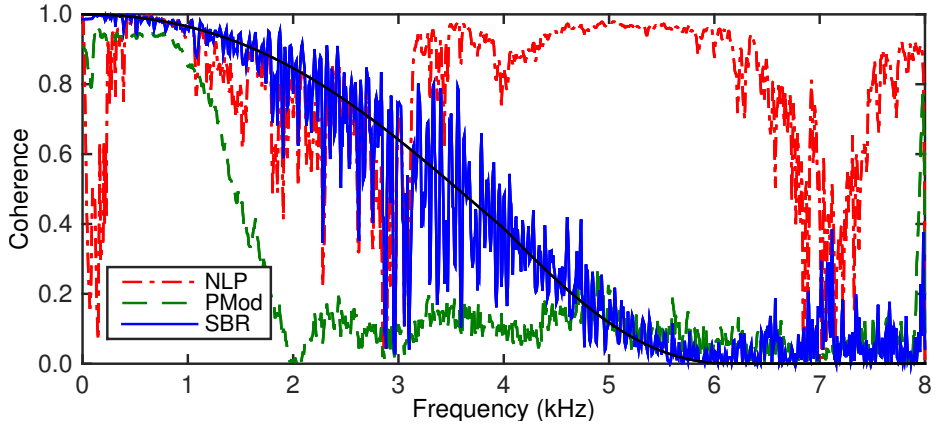
of PMod is low since the magnitude is not changed. Although the LSD of SBR is higher than that of PMod, the distortion of PMod is easily audible through our formal listening test with headphones. PMod heavily distorts the sound image due to the phase modulation process, shifting the image back and forth between the left and the right channels. On the other hand, we observe no noticeable adverse effect to the sound image of SBR. We note that the SSNR, the LSD, and PESQ^{LR} are calculated for individual left and right channels and averaged together, while the PESQ takes a stereo signal as the input and possibly considers the sound image of the input signal. This explains why PMod has high PESQ^{LR}, similar to SBR, and a better LSD but a much lower PESQ than SBR. Overall, decorrelation by resampling with the proposed SBR curve achieves the best processed signal quality as measured by the PESQ, which is high above 4.4.

Figure 31 shows the coherence comparison using different decorrelation procedures, split into two figures for clarity. Figures 32 and 33 show the convergence behavior, where the far-end room impulse response is fixed in Figure 32 while the source location is changed for every 10 seconds in Figure 33. We note that AWGN and NLP are full-band methods without direct control over the coherence in the frequency sub-bands. By comparing AWGN and NLP against no decorrelation, AWGN only slightly decorrelates some high frequency regions, and NLP slightly decorrelates both the low frequency and the high frequency regions. We observe from Figure 32 that NLP and PMod eventually converge to the same lower bound as SBR, while AWGN never converges to that level due to insufficient decorrelation in the low frequency regions.

Although it may appear that the stereo phase modulation procedure [55, Figure 2] is similar to the resampling procedure achieved through the MIDFT in (120), the phase shift introduced by the resampling procedure is an instantaneous linear phase shift in sequential linear increment as opposed to the randomized phase change



(a)



(b)

Figure 31: Coherence comparison with AWGN, NLP, PMod, and SBR. The black solid curve in (b) is the coherence estimated from (135).

in PMod. While both SBR and PMod apply similar concept, i.e., heavy decorrelation in the high frequency sub-bands, the amount of phase shift introduced by the resampling procedure is much smaller compared to PMod since high level of decorrelation can be achieved through a small ΔR. Even though PMod provides some control over the modulation amplitude in each frequency sub-band to adjust the amount of decorrelation, the distortion to the sound image is highly noticeable and significantly degrades the perceptual quality while the resampling procedure provides a much stabler sound image. Furthermore, despite the fact that PMod applies a heavier decorrelation than SBR, there is no performance gain in terms of the convergence rate compared to SBR, as is evident in Figures 31(b), 32, and 33. On the other hand,

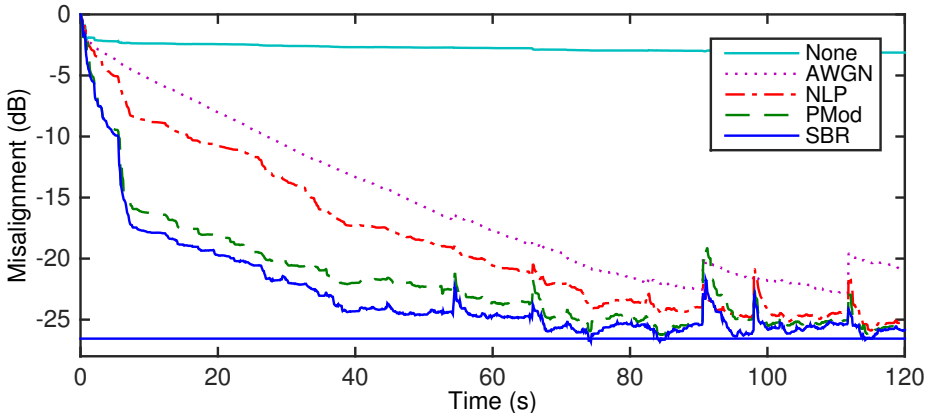


Figure 32: Misalignment comparison with AWGN, NLP, PMod, and SBR. The straight line at the bottom is calculated from (145).

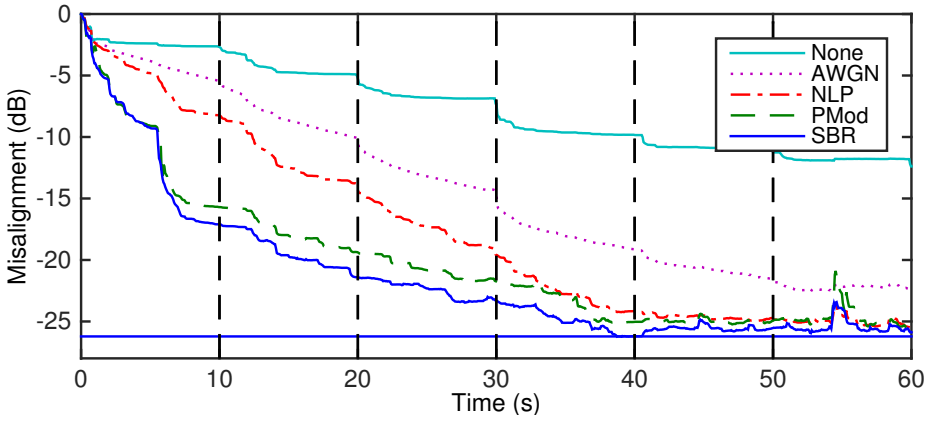


Figure 33: Misalignment comparison. The vertical dotted lines represent the instances when the far-end source location is changed.

by properly designing the proposed SBR curve, the coherence of SBR from the low frequency sub-bands to the high frequency sub-bands decreases smoothly with very low coherence in the high frequency sub-bands and low signal modification in the low frequency sub-bands.

Figure 34 shows the standardized subjective listening results obtained from the MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA) [64] test, where HR stands for "hidden reference" and the anchor is a 3.5 kHz low-pass filtered version of the reference. 3 male and 3 female utterances from the TIMIT database were selected for the test, and the following six sentences were used: "I'd ride the subway but I haven't enough change," "Don't ask me to carry an oily rag like that," "But she

suffered in her off-duty hours," "A muscular abdomen is good for your back," "One could hear a very faint ladylike sigh of relief," and "Pretty soon a woman came along carrying a folded umbrella as a walking stick." The MUSHRA test was performed by 12 (9 of them experienced) listeners using the *Beats Studio*$^{\text{TM}}$ headphone. The audio quality is quantified on a scale from 0 (very bad) to 100 (indistinguishable from the reference). We asked the test subjects to evaluate the audio quality based on both the perceived frequency-domain distortion and the spatial localization accuracy after processing. The mean and the 95% confidence intervals are plotted in Figure 34, and the listeners clearly prefer the audio quality of the proposed SBR. A significant amount of overlap in the 95% confidence intervals of the SBR and the hidden reference indicates that most listeners cannot reliably tell the difference between the SBR and the hidden reference, indicating the superior audio quality after SBR. Both the sound quality and the sound image after SBR remain close to the original reference. Although PMod does not introduce frequency-domain magnitude distortion, the subjective quality score is much lower than SBR due to the constant shifting of the sound image. NLP and AWGN generate too much frequency-domain distortion and receive poor scores. Our experimental comparisons summarized in Table 18 and Figures 32, 33, and 34 clearly demonstrate the superior performance of SBR in terms of not only the convergence rate, but also the processed speech quality.

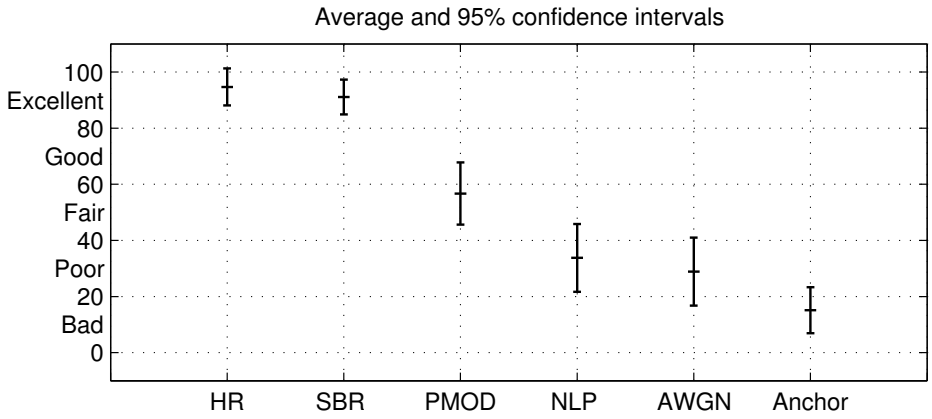


Figure 34: Subjective audio quality comparison from the MUSHRA test.

CHAPTER V

PRACTICAL CONSIDERATIONS FOR VOICE SYSTEMS

We have shown in Chapter 3 the advantage of the system approach for improved acoustic echo cancellation (AEC) and residual echo suppression (RES). We have also presented in Chapter 4 a decorrelation procedure that is designed through the system approach to introduce the least amount of loudspeaker signal distortion while greatly accelerate the convergence rate of a multi-channel acoustic echo cancellation (MCAEC).

However, we have thus far only considered individual component one at a time and optimize the algorithm for limited cases (although inter-modular characteristics have been communicated among various components). When all individual pieces are combined together, the behavior of the whole system can still be very different. Furthermore, the system needs to perform well in a wide variety of acoustic mixing environments with different levels of signal-to-noise ratios (SNRs). Optimal parameter tuning of the system can drastically increase the performance of the whole system. The traditional approach by tuning the components one block at a time using a small set of database quickly becomes inadequate as the system complexity increases with many different components and tuning parameters. Therefore, an alternative automated database generation and tuning approach is presented and discussed in this chapter to complete the essential idea of "a system-based approach" towards echo and distortion management.

For real world applications, the computational complexity of the whole system cannot exceed the processing power of a target platform. We may achieve a good AEC performance through more iterations, but the specific configuration may exceed

the computational budget of the target platform, which can be low for embedded platforms such as a bluetooth loudspeaker. To address this issue, we implement the full algorithm and calculate the actual computational cost for each block as a function of the tuning parameters, e.g., number of iterations. We then formulate the tuning procedure as a constrained optimization problem and show how the computational constraint results in a parameter set that maximizes the voice quality and is still feasible on the target platform.

Finally, we present a robust stereo echo canceler using the decorrelation procedure presented in Chapter 4 and propose an improved robust regularization procedure for the stereo echo canceler.

5.1 Robust Single-Channel Voice System

Let $y[n]$ be the near-end microphone signal, which consists of the near-end speech $s[n]$ and noise $v[n]$ mixed with the acoustic echo $d[n] = h[n] * x[n]$, where $h[n]$ is the impulse response of the system and $x[n]$ is the far-end reference signal. The overall block diagram of the speech enhancement algorithm is shown in Figure 35, which consists of two robust acoustic echo cancellation (RAEC) blocks, a double-talk probability (DTP) estimator, two residual power estimation (RPE) blocks, a noise power estimation (NPE) block, a noise suppression (NS) block and a binary mask.

The system was designed with low computational complexity in mind to allow for deployment on embedded devices. Some of the design choices were made due to the low computational power of the target platform. Detailed descriptions of each block are given in the following sections.

5.1.1 Robust Acoustic Echo Cancellation with Multi-Delay Filter

The AEC algorithms introduced thus far were all block-based AEC algorithms, where the algorithmic delay is equal to the length of the adaptive filter. However, for actual real-time systems, the length of the adaptive filter can be on the order of

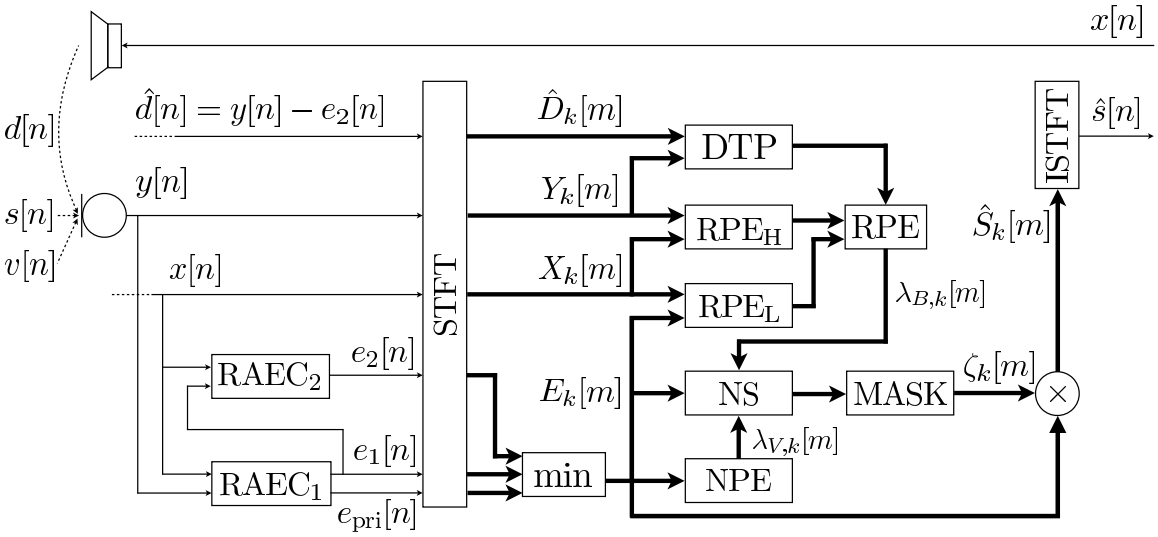


Figure 35: A block diagram of a robust single-channel voice system.

thousand taps at a sampling rate of 16 kHz, leading to a long algorithmic delay on the order of hundred milliseconds. Other delays such as the input/output buffer delay and the transmission delay on top of the algorithmic delay can easily render the algorithm unfeasible for real-time telecommunication. The multi-delay filter (MDF) [105] circumvents this algorithmic delay problem by breaking the adaptive filter into smaller sub-blocks. Given a filter length L, the adaptive filter is broken into M blocks, where the length of each block is N and $MN = L$. The multi-delay adaptive filtering algorithm is summarized in Table 19.

Note that due to the partitioning of the adaptive filter, the gradient constraint has to be applied on each partition for the fully constrained adaptation, resulting in significantly more fast Fourier transform (FFT) operations for each block of incoming signal compared to the regular block frequency-domain adaptive filter. Ultimately, it is a trade-off between computational complexity and algorithmic delay. In the extreme case, when the block size N equals one, the complexity increases to be the same as the time-domain direct convolution while the algorithmic delay reduces to one sample only. The design to balance complexity and algorithmic delay is an important issue when deploying the system on real-time processing platforms.

Table 19: The multi-delay adaptive filter.

Definitions

$$[\mathbf{F}]_{k+1,n+1} = \exp(-j\tfrac{\pi kn}{N}), \quad k,n = 0,\ldots,2N-1$$
$$\mathbf{G}^{01} = \mathbf{F}\begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}$$

Initialization

$$\{\underline{\mathbf{w}}_l[0] = \mathbf{0}_{2N\times 1}, l = 0,\ldots,M-1\}, \quad \{x[n] = 0; n < 0\}, \quad \{\underline{\mathbf{s}}_x[m] = \mathbf{0}_{2N\times 1}; m < 0\}$$
$$MN = L, \quad N > 0, \quad \mu > 0, \quad \epsilon > 0, \quad 0 \ll \beta < 1$$

Spectral estimation

$$\underline{\mathbf{x}}[m] = \mathbf{F}[x[(m-1)N],\ldots,x[(m+1)N-1]]^{\mathrm{T}}$$
$$\underline{\mathbf{s}}_x[m] = \beta\underline{\mathbf{s}}_x[m-1] + (1-\beta)(\underline{\mathbf{x}}[m]\circ\underline{\mathbf{x}}^*[m])$$
$$\underline{\mathbf{n}}[m] = \epsilon\mathbf{1}_{2N\times 1} + \underline{\mathbf{s}}_x[m]$$

Filter adaptation

$$\underline{\mathbf{y}}[m] = \mathbf{F}[\mathbf{0}_{1\times L}, y[mN],\ldots,y[(m+1)N-1]]^{\mathrm{T}}$$
$$\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \mathbf{G}^{01}\left(\sum_{l=0}^{M-1}\underline{\mathbf{w}}_l[m]\circ\underline{\mathbf{x}}[m-l]\right)$$
$$\underline{\mathbf{w}}_l[m+1] = \underline{\mathbf{w}}_l[m] + \mu\mathbf{G}^{10}(\underline{\mathbf{n}}^{\circ(-1)}[m]\circ\underline{\mathbf{e}}[m]\circ\underline{\mathbf{x}}^*[m-l]), \quad l = 0,\ldots,M-1$$

One way to reduce the computational complexity is to utilized the alternate constrained scheme [105] so that the gradient constraint is applied only on one partition per iteration, i.e.,

$$\underline{\mathbf{w}}_l[m+1] = \begin{cases}\underline{\mathbf{w}}_l[m] + \mathbf{G}^{10}(\underline{\boldsymbol{\mu}}[m]\circ\phi(\underline{\mathbf{e}}[m])\circ\underline{\mathbf{x}}^*[m-l]), & m \bmod M = l,\\ \underline{\mathbf{w}}_l[m] + (\underline{\boldsymbol{\mu}}[m]\circ\phi(\underline{\mathbf{e}}[m])\circ\underline{\mathbf{x}}^*[m-l]), & \text{otherwise.}\end{cases} \tag{155}$$

The alternate constrained scheme inevitably slows down the convergence rate but also greatly reduces the computational complexity. For example, consider a typical case at 16 kHz sampling rate where the adaptive filter length is $L = 2048$ and the block size is $N = 256$ (16 ms algorithmic delay) which results in $M = 8$ blocks. If we further assume the number of iteration `numIterations` $= 4$, the fully constrained

adaptation results in 32 FFTs and 32 inverse fast Fourier transforms (IFFTs) while the alternate constrained adaptation requires only 4 FFTs and 4 IFFTs, reducing the number of FFT operations for the gradient constraint by a factor of 8. Even though the computational complexity can be further reduced by using the unconstrained adaptation [83], i.e.,

$$\underline{\mathbf{w}}_l[m+1] = \underline{\mathbf{w}}_l[m] + (\underline{\boldsymbol{\mu}}[m] \circ \phi(\underline{\mathbf{e}}[m]) \circ \underline{\mathbf{x}}^*[m-l]), \tag{156}$$

it is not suggested since the convergence slows down significantly.

Recall from Section 2.3 that the RAEC algorithm contains mainly three elements for the robust update of the adaptive filter even during continuous double talk, i.e., the error recovery nonlinearity (ERN), the noise-robust adaptive step-size, and the iterative adaptation [114–117, 126, 127]. The RAEC algorithm was used in Chapter 3 for the system approach, but the block-based frequency-domain adaptive filter structure was used. Therefore, we formulate the RAEC algorithm with the multi-delay adaptive filter structure such that it is more applicable for real-time systems. The RAEC algorithm that utilizes the MDF structure is shown in Table 20.

Note that we use the assignment operator $\leftarrow$ to facilitate the notation of iterative update. The statistics of the reference signal and the error signal are updated using the information of the latest block, i.e., $\underline{\mathbf{x}}[m]$ and $\underline{\mathbf{e}}[m]$, respectively, and the error signal of the latest block is used to update all adaptive filter blocks. Since the ERN effectively limits the amplitude of the error signal, the convergence rate can potentially be slowed down when there is no near-end noise. The iteration helps in such a case to regain the convergence speed at the expense of higher computational complexity.

To reduce the complexity of the RAEC algorithm, we elect to use the alternate constrained weight update to reduce the number of FFTs and IFFTs. Certain operations of the hybrid RAEC in Chapter 3 can be costly and therefore a cascaded structure similar to the system approach of [127] is used: the output of the first RAEC is fed to the input of the second RAEC, which is different from the original system

Table 20: The RAEC algorithm with MDF.

Definitions

$$[\mathbf{F}]_{k+1,n+1} = \exp(-j\tfrac{\pi kn}{N}), \quad k,n = 0,\ldots,2N-1$$
$$\mathbf{G}^{01} = \mathbf{F}\begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}$$
$$\phi(\underline{\mathbf{e}}) \equiv [\phi(E_0),\ldots,\phi(E_{2N-1})]^{\mathrm{T}}$$

Initialization

$$\{\underline{\mathbf{w}}_l = \mathbf{0}_{2N\times 1}, l = 0,\ldots,M-1\}, \quad \{x[n]=0; n<0\}, \quad \underline{\mathbf{s}}_x = \mathbf{0}_{2N\times 1}, \quad \underline{\mathbf{s}}_e = \mathbf{0}_{2N\times 1}$$
$$MN = L, \quad N>0, \quad \mu>0, \quad \gamma>0, \quad \epsilon>0, \quad 0 \ll \beta < 1$$

Iterative filter adaptation

$$\underline{\mathbf{x}}[m] = \mathbf{F}[x[(m-1)N],\ldots,x[(m+1)N-1]]^{\mathrm{T}}$$
$$\underline{\mathbf{y}}[m] = \mathbf{F}[\mathbf{0}_{1\times L}, y[mN],\ldots,y[(m+1)N-1]]^{\mathrm{T}}$$

`for` $i :=$ `1 to numIterations`

$$\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \mathbf{G}^{01}\left(\sum_{l=0}^{M-1}\underline{\mathbf{w}}_l \circ \underline{\mathbf{x}}[m-l]\right)$$
$$\underline{\mathbf{s}}_x \leftarrow \beta\underline{\mathbf{s}}_x + (1-\beta)(\underline{\mathbf{x}}[m]\circ\underline{\mathbf{x}}^*[m])$$
$$\underline{\mathbf{s}}_e \leftarrow \beta\underline{\mathbf{s}}_e + (1-\beta)(\underline{\mathbf{e}}[m]\circ\underline{\mathbf{e}}^*[m])$$
$$\phi(E_k[m]) = \begin{cases}\sqrt{S_{ee}[k]}\, e^{j\angle E_k[m]}, & |E_k[m]| > \sqrt{S_{ee}[k]}\\ E_k[m], & \text{otherwise}\end{cases}$$
$$\underline{\boldsymbol{\mu}}[m] = \mu\underline{\mathbf{s}}_x \circ [(\underline{\mathbf{s}}_x)^{\circ 2} + \gamma(\underline{\mathbf{s}}_e)^{\circ 2} + \epsilon\mathbf{1}_{2N\times 1}]^{\circ(-1)}$$
$$\underline{\mathbf{w}}_l \leftarrow \begin{cases}\underline{\mathbf{w}}_l + \mathbf{G}^{10}(\underline{\boldsymbol{\mu}}[m]\circ\phi(\underline{\mathbf{e}}[m])\circ\underline{\mathbf{x}}^*[m-l]), & m \bmod M = l\\ \underline{\mathbf{w}}_l + (\underline{\boldsymbol{\mu}}[m]\circ\phi(\underline{\mathbf{e}}[m])\circ\underline{\mathbf{x}}^*[m-l]), & \text{otherwise}\end{cases}$$

`end for`

approach where the input to the second RAEC is still the microphone signal (a parallel structure instead of the cascaded structure used in this section). Note that the signal $e_{\mathrm{pri}}[n]$ in Figure 35 is the error signal before the adaptive filter update, whereas the signals $e_1[n]$ and $e_2[n]$ are the error signals after the filter update. The tuning parameters for each of the RAECs consist of the frame size N_{RAEC}, the number of

partitioned blocks M_{RAEC}, the number of iterations N_{iter}, the step-size μ_{RAEC}, the tuning parameter γ_{RAEC} for the robust adaptive step-size, and the smoothing factor α_{RAEC} for the power spectral density estimation.

While some tuning parameters such as the step-size may affect only the system performance, other tuning parameters such as the frame size and the number of partitioned blocks may directly affect the computational complexity and/or delay. Even though we may manually optimize the AEC to balance the complexity and echo cancellation performance, the overall system complexity and performance is not guaranteed to be optimal, i.e., locally optimized parameters for each block may not translate to globally optimal performance when all the components are connected. Therefore, we delay the discussion of system performance until Section 5.2, where we introduce the automated tuning framework with computational complexity constraint.

5.1.2 Residual Echo Power Estimator

Since the AEC cannot cancel all the echo signal due to modeling mismatch, further enhancement from the RES is required to improve the voice quality. A coherence based method similar to [28, 45] is used for the RPE, and a modified version of the DTP estimator similar to [110] is used for a more accurate estimate of the residual echo power. As shown in Figure 35, the DTP estimator differs from that in [110] since the coherence is calculated between the RAEC estimated echo signal $\hat{d}$ and the microphone signal y rather than between the loudspeaker signal x and the microphone signal y. This is possible since the estimated echo signal $\hat{d}$ can be reliably obtained even during double talk due to the *robust* echo path tracking performance of the RAEC.

As shown in (164), the DTP estimator differs from that in [110] since the coherence is calculated between the RAEC estimated echo signal $\hat{d}$ and the microphone signal y rather than between the loudspeaker signal x and the microphone signal y. The

traditional way to estimate the coherence in [110] is to use the following statistical parameters:

$$\Phi_{XX,k}[m] = \alpha_{\mathrm{DTP}}\Phi_{XX,k}[m-1] + (1-\alpha_{\mathrm{DTP}})|X_k[m]|^2, \tag{157}$$

$$\Phi_{YY,k}[m] = \alpha_{\mathrm{DTP}}\Phi_{YY,k}[m-1] + (1-\alpha_{\mathrm{DTP}})|Y_k[m]|^2, \tag{158}$$

$$\Phi_{XY,k}[m] = \alpha_{\mathrm{DTP}}\Phi_{XY,k}[m-1] + (1-\alpha_{\mathrm{DTP}})X_k^*[m]Y_k[m], \tag{159}$$

$$\rho_{\mathrm{old},k}[m] = \frac{|\Phi_{XY,k}[m]|^2}{\Phi_{XX,k}[m]\Phi_{YY,k}[m]}. \tag{160}$$

The calculation of the coherence using:

$$\Phi_{\hat{D}\hat{D},k}[m] = \alpha_{\mathrm{DTP}}\Phi_{\hat{D}\hat{D},k}[m-1] + (1-\alpha_{\mathrm{DTP}})|\hat{D}_k[m]|^2 \tag{161}$$

$$\Phi_{YY,k}[m] = \alpha_{\mathrm{DTP}}\Phi_{YY,k}[m-1] + (1-\alpha_{\mathrm{DTP}})|Y_k[m]|^2 \tag{162}$$

$$\Phi_{\hat{D}Y,k}[m] = \alpha_{\mathrm{DTP}}\Phi_{\hat{D}Y,k}[m-1] + (1-\alpha_{\mathrm{DTP}})\hat{D}_k^*[m]Y_k[m] \tag{163}$$

$$\rho_{\mathrm{new},k}[m] = \frac{|\Phi_{\hat{D}Y,k}[m]|^2}{\Phi_{\hat{D}\hat{D},k}[m]\Phi_{YY,k}[m]} \tag{164}$$

is possible since the estimated echo signal $\hat{d}$ can be reliably obtained even during double talk due to the *robust* echo path tracking performance of the RAEC. Therefore, the coherence measure ρ_k can be estimated more reliably based on the estimated echo $\hat{d}$ and the microphone signal y.

We propose to estimate the residual echo power by utilizing the output of the double talk probability estimator. Ideally, when the double-talk probability is high, the level of residual echo power estimate should be low so as to not distort the near-end speech when suppressing the residual echo. On the other hand, when the double-talk probability is low, the level of residual echo power estimate should be high to suppress as much residual echo as possible. The high level residual echo power $\lambda_{B_{\mathrm{H}},k}$

$$\lambda_{B_{\mathrm{H}},k}[m] = \left| \frac{\Phi^{\mathrm{T}}_{\mathbf{X}E,k}[m]\mathbf{X}_{\mathrm{H},k}[m]}{\Phi_{X_{\mathrm{H}},k}[m]} \right|^2 \tag{165}$$

is estimated based on the coherence of the microphone signal Y_k and the reference

signal X_k, while the low level residual echo power $\lambda_{B_{\mathrm{L}},k}$

$$\lambda_{B_{\mathrm{L}},k}[m] = \left| \frac{\Phi^{\mathrm{T}}_{\mathbf{X}E,k}[m]\mathbf{X}_{\mathrm{L},k}[m]}{\Phi_{X_{\mathrm{L}},k}[m]} \right|^2 \tag{166}$$

is estimated based on the coherence of the error signal E_k and the reference signal X_k. Finally, the residual echo power $\lambda_{B,k}$ is estimated by utilizing the double-talk probability estimate P_k^{DT}

$$P_k^{\mathrm{DT}}[m] = \frac{\Lambda^{\mathrm{DT}}[m]}{1+\Lambda^{\mathrm{DT}}[m]} \frac{\Lambda_k^{\mathrm{DT}}[m]}{1+\Lambda_k^{\mathrm{DT}}[m]} \tag{167}$$

obtained from DTP to combine $\lambda_{B_{\mathrm{H}},k}$ and $\lambda_{B_{\mathrm{L}},k}$:

$$\lambda_{B,k}[m] = (1 - P_k^{\mathrm{DT}}[m])\lambda_{B_{\mathrm{H}},k}[m] + P_k^{\mathrm{DT}}[m]\lambda_{B_{\mathrm{L}},k}[m]. \tag{168}$$

The idea is to utilize the low level residual echo power estimate during double-talk so that the speech distortion can be reduced, while a higher level residual echo power estimate can be used when there is no double-talk.

The tuning parameters for the DTP consists of the transition probabilities a_{01}, a_{10}, b_{01}, and b_{10}, the smoothing factors α_{DTP} and β_{DTP}, the frequency bin range $[k_{\mathrm{begin}}, k_{\mathrm{end}}]$, the frame duration T_{DTP}, and the adaptation time constants $\tau^{(0)}$ and $\tau^{(1)}$, where $\{\cdot\}^{(0)}$ is for the statistical parameters corresponding to the non-double-talk state and $\{\cdot\}^{(1)}$ is for that of the double-talk state. The tuning parameters for the RPE consist of the numbers of partitions $M_{\mathrm{RPE_H}}$ and $M_{\mathrm{RPE_L}}$ to calculate the coherence and the smoothing factors $\alpha_{\mathrm{RPE_H}}$ and $\alpha_{\mathrm{RPE_L}}$ for the power spectral density estimation. The RPE with the DTP estimator is summarized in Table 21.

5.1.3 Residual Echo and Noise Suppressor

The low complexity minimum mean squared error (MMSE) noise power estimator [41] that implicitly accounts for the speech presence probability (SPP) is used for the NPE. The MMSE estimation of a noisy periodogram under speech presence uncertainty results in

$$\mathrm{E}\{\lambda_{V,k}[m]|E_k[m]\} = P(H_1|E_k[m])\lambda_{V,k}[m] + P(H_0|E_k[m])|E_k[m]|^2, \tag{169}$$

Table 21: Double-talk probability and residual power estimator.

Double-talk probability estimation

$$
\begin{aligned}
\Phi_{\hat{D}\hat{D},k}[m] &= \alpha_{\mathrm{DTP}}\Phi_{\hat{D}\hat{D},k}[m-1] + (1-\alpha_{\mathrm{DTP}})|\hat{D}_k[m]|^2 \\
\Phi_{YY,k}[m] &= \alpha_{\mathrm{DTP}}\Phi_{YY,k}[m-1] + (1-\alpha_{\mathrm{DTP}})|Y_k[m]|^2 \\
\Phi_{\hat{D}Y,k}[m] &= \alpha_{\mathrm{DTP}}\Phi_{\hat{D}Y,k}[m-1] + (1-\alpha_{\mathrm{DTP}})\hat{D}_k^*[m]Y_k[m] \\
\rho_k[m] &= |\Phi_{\hat{D}Y,k}[m]|^2/(\Phi_{\hat{D}\hat{D},k}[m]\Phi_{YY,k}[m]) \\
\Lambda_k[m] &= \sqrt{\frac{\lambda_k^{(0)}[m-1]}{\lambda_k^{(1)}[m-1]}}\frac{\exp\{(\rho_k[m]-\bar{\rho}_k^{(0)}[m-1])^2/\lambda_k^{(0)}[m-1]\}}{\exp\{(\rho_k[m]-\bar{\rho}_k^{(1)}[m-1])^2/\lambda_k^{(1)}[m-1]\}} \\
\Lambda_k^{\mathrm{DT}}[m] &= \frac{a_{01}+(1-a_{10})\Lambda_k^{\mathrm{DT}}[m-1]}{(1-a_{01})+a_{10}\Lambda_k^{\mathrm{DT}}[m-1]}\Lambda_k[m] \\
\Lambda^{\mathrm{GM}}[m] &= \exp\left\{\frac{1}{k_{\mathrm{end}}-k_{\mathrm{begin}}+1}\sum\nolimits_{k=k_{\mathrm{begin}}}^{k_{\mathrm{end}}}\log(\Lambda_k^{\mathrm{DT}}[m])\right\} \\
\Lambda^{\mathrm{AM}}[m] &= \frac{1}{k_{\mathrm{end}}-k_{\mathrm{begin}}+1}\sum\nolimits_{k=k_{\mathrm{begin}}}^{k_{\mathrm{end}}}\Lambda_k^{\mathrm{DT}}[m] \\
\Lambda[m] &= \beta_{\mathrm{DTP}}\Lambda^{\mathrm{GM}}[m] + (1-\beta_{\mathrm{DTP}})\Lambda^{\mathrm{AM}}[m] \\
\Lambda^{\mathrm{DT}}[m] &= \frac{b_{01}+(1-b_{10})\Lambda^{\mathrm{DT}}[m-1]}{(1-b_{01})+b_{10}\Lambda^{\mathrm{DT}}[m-1]}\Lambda[m] \\
P_k^{\mathrm{DT}}[m] &= \frac{\Lambda^{\mathrm{DT}}[m]}{1+\Lambda^{\mathrm{DT}}[m]}\frac{\Lambda_k^{\mathrm{DT}}[m]}{1+\Lambda_k^{\mathrm{DT}}[m]} \\
\alpha^{(0)} &= (1-P_k^{\mathrm{DT}}[m])T_{\mathrm{DTP}}/\tau^{(0)} \\
\bar{\rho}_k^{(0)}[m] &= (1-\alpha^{(0)})\bar{\rho}_k^{(0)}[m-1] + \alpha^{(0)}\rho_k[m] \\
\lambda_k^{(0)}[m] &= (1-\alpha^{(0)})\lambda_k^{(0)}[m-1] + \alpha^{(0)}(\rho_k[m]-\bar{\rho}_k^{(0)}[m])^2 \\
\alpha^{(1)} &= P_k^{\mathrm{DT}}[m]T_{\mathrm{DTP}}/\tau^{(1)} \\
\bar{\rho}_k^{(1)}[m] &= (1-\alpha^{(1)})\bar{\rho}_k^{(1)}[m-1] + \alpha^{(1)}\rho_k[m] \\
\lambda_k^{(1)}[m] &= (1-\alpha^{(1)})\lambda_k^{(1)}[m-1] + \alpha^{(1)}(\rho_k[m]-\bar{\rho}_k^{(1)}[m])^2
\end{aligned}
$$

Residual echo power estimation

$$
\begin{aligned}
\mathbf{X}_{\mathrm{H},k}[m] &= [X_k[m],\ldots,X_k[m-M_{\mathrm{RPE_H}}+1]]^{\mathrm{T}} \\
\Phi_{\mathbf{X}E,k}[m] &= \alpha_{\mathrm{RPE_H}}\Phi_{\mathbf{X}E,k}[m-1] + (1-\alpha_{\mathrm{RPE_H}})(\mathbf{X}_{\mathrm{H},k})^*E_k \\
\Phi_{X_{\mathrm{H}},k}[m] &= \alpha_{\mathrm{RPE_H}}\Phi_{X_{\mathrm{H}},k}[m-1] + (1-\alpha_{\mathrm{RPE_H}})|X_k[m]|^2 \\
\lambda_{B_{\mathrm{H}},k}[m] &= |\Phi_{\mathbf{X}E,k}^{\mathrm{T}}[m]\mathbf{X}_{\mathrm{H},k}[m]/\Phi_{X_{\mathrm{H}},k}[m]|^2 \\
\mathbf{X}_{\mathrm{L},k}[m] &= [X_k[m],\ldots,X_k[m-M_{\mathrm{RPE_L}}+1]]^{\mathrm{T}} \\
\Phi_{\mathbf{X}E,k}[m] &= \alpha_{\mathrm{RPE_L}}\Phi_{\mathbf{X}E,k}[m-1] + (1-\alpha_{\mathrm{RPE_L}})(\mathbf{X}_{\mathrm{L},k})^*E_k \\
\Phi_{X_{\mathrm{L}},k}[m] &= \alpha_{\mathrm{RPE_L}}\Phi_{X_{\mathrm{L}},k}[m-1] + (1-\alpha_{\mathrm{RPE_L}})|X_k[m]|^2 \\
\lambda_{B_{\mathrm{L}},k}[m] &= |\Phi_{\mathbf{X}E,k}^{\mathrm{T}}[m]\mathbf{X}_{\mathrm{L},k}[m]/\Phi_{X_{\mathrm{L}},k}[m]|^2 \\
\lambda_{B,k}[m] &= (1-[m]P_k^{\mathrm{DT}}[m])\lambda_{B_{\mathrm{H}},k}[m] + P_k^{\mathrm{DT}}[m]\lambda_{B_{\mathrm{L}},k}[m]
\end{aligned}
$$

where the *a posteriori* SPP is calculated by

$$P(H_1|E_k[m]) = \left[1 + (1+\xi_{H_1})\exp\left(- \frac{|E_k[m]|^2}{\lambda_{V,k}[m-1]} \frac{\xi_{H_1}}{1+\xi_{H_1}} \right)\right]^{-1}. \tag{170}$$

The noise power spectral density is then updated by

$$\lambda_{V,k}[m] = \alpha_{\mathrm{NPE}}\lambda_{V,k}[m-1] + (1-\alpha_{\mathrm{NPE}})\mathrm{E}\{\lambda_{V,k}[m]|E_k[m]\}. \tag{171}$$

To avoid stagnation due to an underestimated noise power, a smoothing is performed

$$\bar{P}_k[m] = \alpha_P \bar{P}_k[m-1] + (1-\alpha_P)P(H_1|E_k[m]), \tag{172}$$

and the following ad-hoc procedure is used for the update:

$$P(H_1|E_k[m]) = \begin{cases} \min\{P(H_1|E_k[m]), P_{\mathrm{TH}}\}, & \bar{P}_k[m] > P_{\mathrm{TH}}, \\ P(H_1|E_k[m]), & \text{otherwise.} \end{cases} \tag{173}$$

We combine RPE and NPE for residual echo and noise suppression using a single noise suppressor, as shown in Figure 35. The Ephraim and Malah log-spectral amplitude (LSA) estimator [30] is used for the combined residual echo and noise suppression:

$$G_k^{\mathrm{LSA}}[m] = \frac{\xi_k[m]}{1+\xi_k[m]} \exp\left(\frac{1}{2} \int_{\frac{\xi_k[m]\gamma_k[m]}{1+\xi_k[m]}}^{\infty} \frac{e^{-t}}{t}\, \mathrm{d}t \right). \tag{174}$$

The estimation of the *a priori* SNR ξ_k is done using the decision-directed (DD) approach [29]:

$$\xi_k[m] = \alpha_{\mathrm{DD}} \frac{|\hat{S}_k[m-1]|^2}{\lambda_{V,k}[m] + \lambda_{B,k}[m]} + (1-\alpha_{\mathrm{DD}})\max\{\gamma_k[m]-1, 0\}, \tag{175}$$

where

$$\gamma_k[m] = \frac{\lambda_{E,k}[m]}{\lambda_{V,k}[m] + \lambda_{B,k}[m]} \tag{176}$$

and $\lambda_{E,k}$, $\lambda_{V,k}$, and $\lambda_{B,k}$ are the residual error signal power, the noise power, and residual echo power respectively. To further reduce the musical noise, the suppression gain is limited to a certain minimum value $G_{\min}$:

$$\hat{S}_k[m] = ((1-G_{\min})G_k^{\mathrm{LSA}}[m] + G_{\min})E_k[m]. \tag{177}$$

The tuning parameters of the NPE consist of the fixed *a priori* SNR ξ_{H_1}, the threshold P_{TH}, and the smoothing factors α_P and α_{NPE} The tuning parameters of the the NS consist of the smoothing factor for the SNR estimator α_{DD} and the minimum gain $G_{\min}$. The residual echo and noise suppressor algorithm is summarized in Table 22.

Table 22: Residual echo and noise suppressor.

Noise power estimation

$$
\begin{aligned}
P(H_1|E_k[m]) &= \left[1+(1+\xi_{H_1})\exp\left(-\frac{|E_k[m]|^2}{\lambda_{V,k}[m-1]}\frac{\xi_{H_1}}{1+\xi_{H_1}}\right)\right]^{-1} \\
\bar{P}_k[m] &= \alpha_P \bar{P}_k[m] + (1-\alpha_P)P(H_1|E_k[m]) \\
P(H_1|E_k[m]) &= \begin{cases} \min\{P(H_1|E_k[m]), P_{\mathrm{TH}}\}, & \bar{P}_k[m] > P_{\mathrm{TH}} \\ P(H_1|E_k[m]), & \text{otherwise} \end{cases} \\
\mathrm{E}\{\lambda_{V,k}[m]|E_k[m]\} &= P(H_1|E_k[m])\lambda_{V,k}[m] + P(H_0|E_k[m])|E_k[m]|^2 \\
\lambda_{V,k}[m] &= \alpha_{\mathrm{NPE}}\lambda_{V,k}[m-1] + (1-\alpha_{\mathrm{NPE}})\mathrm{E}\{\lambda_{V,k}[m]|E_k[m]\}
\end{aligned}
$$

Noise suppressor

$$
\begin{aligned}
\xi_k[m] &= \alpha_{\mathrm{DD}}\frac{|\hat{S}_k[m-1]|^2}{\lambda_{V,k}[m]+\lambda_{B,k}[m]} + (1-\alpha_{\mathrm{DD}})\max\{\gamma_k[m]-1,0\} \\
\gamma_k[m] &= \lambda_{E,k}[m]/(\lambda_{V,k}[m]+\lambda_{B,k}[m]) \\
G_k^{\mathrm{LSA}}[m] &= \frac{\xi_k[m]}{1+\xi_k[m]}\exp\left(\frac{1}{2}\int_{\frac{\xi_k[m]\gamma_k[m]}{1+\xi_k[m]}}^{\infty}\frac{e^{-t}}{t}\,\mathrm{d}t\right) \\
\hat{S}_k[m] &= ((1-G_{\min})G_k^{\mathrm{LSA}}[m] + G_{\min})E_k[m]
\end{aligned}
$$

5.1.4 Quasi-Binary Mask for Speech Recognition

It has been recently shown that the speech recognition accuracy in noisy condition can be greatly improved by direct binary masking [52] when compared to marginalization [21] or spectral reconstruction [95]. Given our application scenario, we propose to combine the direct masking approach, particularly effective at low overall SNRs, with the NS output mask G_k^{LSA}, as shown in Figure 35. In particular, we exploit the

estimated bin-based *a priori* SNR ξ_k to determine the type of masking to be applied to the spectrum. However, given that an accurate estimation of the binary mask is very difficult for very low SNRs, we elect to use the LSA estimated gain for those cases. Our masking then becomes:

$$\zeta_k[m] = \begin{cases} (1 - G_{\min})G_k^{\text{LSA}}[m] + G_{\min}, & \xi_k[m] \leq \theta_1, \\ \frac{\alpha}{2}, & \theta_1 < \xi_k[m] < \theta_2, \\ \frac{2+\alpha}{2}, & \xi_k[m] \geq \theta_2, \end{cases}$$

where $G_{\min}$ is the minimum suppression gain [42], and the output is then:

$$\hat{S}_k[m] = \zeta_k[m]E_k[m]. \tag{178}$$

The tuning parameters for the direct masking consist of the SNR thresholds θ_1 and θ_2, the tuning parameter α, and a binary variable b_m that chooses the type of masking applied (based on the LSA gain (177) or the *quasi*-binary mask).

Although all of the components presented above have been individually studied and carefully tuned in the past, integrating all the components presents an enormous challenge for tuning of the whole system. Due to the number of tuning parameters and the possible interaction between different components, parameter tuning becomes nontrivial, and a system tuned with suboptimal parameters may result in huge degradation of the overall system performance. The tuning problem is further complicated by the fact that some of the tuning parameters affects both the system performance, e.g., echo return loss enhancement (ERLE), and the computational complexity. For example, increasing the number of iterations of the AEC directly improves the convergence speed at the expense of increasing the computational complexity. In the next section, we formalize the tuning problem as a nonlinear optimization problem with computational complexity constraint, and compare the manually tuned system to the proposed tuning methodology.

5.2 Automated Tuning of the Single-Channel Voice System

Very little work has been done to formalize the tuning problem in speech enhancement (SE) systems, notably [111], due to the combinatorial nature of the problem and the related optimization criteria that rely on the fuzzy concept of *perceptually better quality* [42]. To get around the subjective and combinatorial nature of the design and tuning problem, locally optimal or near-optimal solutions are found by considering one component of the system at a time, and the concept of perceived quality is approximated by measures that are easy to describe mathematically, e.g., the mean squared error (MSE) or maximum likelihood (ML) [112]. However, it is well known that these types of measures, as well as the assumptions behind them, are hardly related to the auditory system [18], making the tuned solution suboptimal.

Several methods have been proposed to objectively measure the perceived quality of speech signals [60]. The mean opinion score (MOS) is the current standardized measure which compares a high quality fixed reference to its degraded version and ranks the result from "inaudible" to "very annoying" on a five-point scale [69]. This score can be calculated using automated techniques that mimic the human hearing process [88]. The most commonly used method is the Performance Evaluation of Speech Quality (PESQ) [71], but its scope is limited to speech codecs evaluation. A new model called Perceptual Objective Listening Quality Assessment (POLQA) [72] addresses many of the issues and limitations of PESQ and produces reliable scores for evaluating SE algorithms. When SE systems are used as a pre-processor for automatic speech recognition (ASR), the objective of the algorithmic design is to maximize the speech recognition accuracy [73]. While enhancement methods which facilitate proper adjustments of model parameters have been shown to better account for the mismatch between the training condition and the application scenario [74], methods relying on fixed acoustic models using the hidden Markov models (HMMs) are still the most common methods for limited-vocabulary recognition on embedded

systems [23]. Therefore, these methods rely heavily on the SE algorithms to enhance the speech signals before feature extraction to match the training condition of the ASR [78]. Accurate assessment of the ASR reliability is still a matter of debate since they are heavily application and context dependent [31]. However, for embedded systems, the phone accuracy rate (PAR), or at a higher semantic level the word accuracy rate (WAR), is generally appropriate as a performance measure for the ASR.

However, the objective of maximizing the perceptual quality or the speech recognition accuracy often contradicts the computational constraints imposed by the target platform. While *profiling* each component of an SE system during development is a good practice to avoid overly complex solutions, the *tuning* of the system is often done at an advanced stage of the development and may influence the computational complexity dramatically. During development and prototyping, a commercially viable SE system must also take into account the constraints of the target platform [80]. For audio related applications, field-programmable gate arrays (FPGAs) [87] and dedicated digital signal processors (DSPs) are the most common choices since they generally have lower cost, lower latency, and lower energy consumption [103]. Meeting the computational budget of the target hardware, commonly measured in terms of million cycles per second (MCPS), is generally a dictating condition [93]. The computational complexity of an algorithm is calculated by counting the number of basic mathematical operations, e.g., multiplications, additions, or multiply-accumulations (MACs), as well as the usage of pre-defined, highly-optimized subroutines already embedded in the processor, e.g., the FFTs [79].

Besides the optimization criteria, constructing a comprehensive database that covers all possible scenarios is also essential to developing an effective SE algorithm, and recent works have focused on providing a common framework to test and evaluate SE algorithms, e.g., for noise suppression [59] or dereverberation [76]. However, to

the authors' knowledge, there is currently no database for evaluating SE algorithms in full-duplex communication, which is the target of our system. Thus, a full-duplex communication database is often "handmade" but tailors to only a few scenarios.

In this section, we propose a formal procedure for tuning the parameters of an SE system for hands-free devices [42–44]. The tuning problem is casted as an optimization problem where the cost function is a perceptual objective measure or the back-end recognizer accuracy and the optimization variables are the parameters of the SE chain, and a genetic algorithm is used to determine the global solution. The idea of automatic tuning of system is consistent with the original notion of integrated optimization in the system-based approach, although the relationship between parameter values and objective functions may be quite nonlinear and difficult to control. Similar ideas were used in [111] and in [121], to tune the parameters of a noise reduction system and the parameters of a ASR back-end, respectively. A nonlinear penalty function accounting for the computational complexity is introduced in the optimization framework to account for the computational complexity. For this purpose, a large multi-condition database is automatically generated by considering the characteristics of human conversational speech. The database encompasses various key factors including room impulse responses (RIRs), noise types, speakers, echo return losses, and SNRs, to model a real full-duplex communication. We first compare different objective perceptual measures as optimization criteria and perform a subjective listening test on the different outputs obtained using an unconstrained optimization. The system is then optimized for either full-duplex communications or an ASR front-end with the computational complexity constraint specified in terms of MCPS.

5.2.1 Computational Complexity of the Voice System

On the system level tuning, all tuning parameters are listed as follows. The tuning parameters for each of the RAECs consist of the frame size N_{RAEC}, the number of partitioned blocks M_{RAEC}, the number of iterations N_{iter}, the step-size μ_{RAEC}, the tuning parameter γ_{RAEC} for the robust adaptive step-size, and the smoothing factor α_{RAEC} for the power spectral density estimation. The tuning parameters for the DTP consists of the transition probabilities a_{01}, a_{10}, b_{01}, and b_{10}, the smoothing factors α_{DTP} and β_{DTP}, the frequency bin range $[k_{\mathrm{begin}}, k_{\mathrm{end}}]$, the frame duration T_{DTP}, and the adaptation time constants τ. The tuning parameters for the RPE consist of the numbers of partitions $M_{\mathrm{RPE_H}}$ and $M_{\mathrm{RPE_L}}$ to calculate the coherence and the smoothing factors $\alpha_{\mathrm{RPE_H}}$ and $\alpha_{\mathrm{RPE_L}}$ for the power spectral density estimation. The tuning parameters of the NPE consist of the fixed *a priori* SNR ξ_{H_1}, the threshold P_{TH}, and the smoothing factors α_P and α_{NPE} The tuning parameters of the the NS consist of the smoothing factor for the SNR estimator α_{DD} and the minimum gain G_{min}. The tuning parameters for the direct masking consist of the SNR thresholds θ_1 and θ_2, the tuning parameter α, and a binary variable b_m that chooses between the LSA gain [29] or the *quasi*-binary.

Table 23 shows the computational complexity per sample for each block, where "mply" stands for multiplication, "add" stands for addition, "sqrt" stands for square root, "if-else" stands for the if-else statement, "div" stands for division, "log" stands for the logarithm function, "exp" stands for the exponential function, "MAC" stands for multiply-accumulation, "cplx" stands for complex number, and "pwrSpectr" stands for the square of the magnitude of a complex number. Eventually, the actual complexity is platform dependent, but each of the fundamental operations, such as the FFT, can be estimated in terms of DSP cycles, which in turn allows us to estimate the computation on an actual platform in terms of MCPS. Note that $\mathrm{FFT_{RAEC}}$ and $\mathrm{FFT_{STFT}}$ represent the FFT cost per sample by dividing the FFT cost by its block

Table 23: The computational complexity per sample for each block.

$$
\begin{aligned}
C_{\text{RAEC}} &= (3N_{\text{iter}}+2)\text{-FFT}_{\text{RAEC}} + (5N_{\text{iter}}+3)\text{-mply} + (3N_{\text{iter}}+1)\text{-MAC} \\
&\quad + (2N_{\text{iter}}+1)\text{-cplx-pwrSpectr} + (2N_{\text{iter}}+1)M_{\text{RAEC}}\text{-cplx-mply} \\
&\quad + N_{\text{iter}}(M_{\text{RAEC}}+1)\text{-add} + N_{\text{iter}}\text{-sqrt} + 2N_{\text{iter}}\text{-div} + N_{\text{iter}}\text{-if-else} \\
&\quad + N_{\text{iter}}M_{\text{RAEC}}\text{-real-cplx-mply} \\
C_{\text{STFT}} &= 2\text{-mply} + \text{FFT}_{\text{STFT}} \\
C_{\text{DTP}} &= 3\text{-cplx-pwrSpectr} + 18\text{-mply} + 12\text{-MAC} + 1\text{-cplx-mply} + 6\text{-div} \\
&\quad + 9\text{-add} + 1\text{-exp} + 1\text{-sqrt} + 1\text{-log} \\
C_{\text{RPE}} &= 1\text{-cplx-pwrSpectr} + 4\text{-mply} + 3\text{-MAC} + (M_{\text{RPE}}+1)\text{-cplx-mply} \\
&\quad + (M_{\text{RPE}}+1)\text{-add} + 1\text{-div} \\
C_{\text{NPE}} &= 1\text{-cplx-pwrSpectr} + 3\text{-div} + 3\text{-add} + 5\text{-mply} + 1\text{-exp} + 3\text{-MAC} \\
&\quad + 2\text{-if-else} \\
C_{\text{NS}} &= 2\text{-cplx-pwrSpectr} + 2\text{-add} + 1\text{-if-else} + 3\text{-mply} + 2\text{-MAC} + 3\text{-div}
\end{aligned}
$$

size. Also note that some of the tuning parameters, such as the number of partitioned blocks M_{RAEC} and M_{RPE}, the $2N_{\text{RAEC}}$-point FFT of the RAEC, the N_{STFT}-point FFT of the short-time Fourier transform (STFT) block, and the number of iterations N_{iter}, will influence directly the complexity. Given the computational complexity of each block, the total computational complexity in terms of MCPS is given by

$$
\begin{aligned}
\text{C}(\mathbf{p}) = (&C_{\text{RAEC}_1} + C_{\text{RAEC}_2} + 7C_{\text{STFT}} + C_{\text{DTP}} \\
&+ C_{\text{RPE}_{\text{H}}} + C_{\text{RPE}_{\text{L}}} + C_{\text{NPE}} + C_{\text{NS}})\frac{f_s}{10^6}\ [\text{MCPS}],
\end{aligned} \tag{179}
$$

where $\mathbf{p}$ is the vector of optimization parameters and f_s is the sampling rate. Additionally, there is an on-off flag to either turn on or off the second RAEC block to determine whether using the cascaded structure of two RAEC blocks or running only one RAEC block for a higher number of iterations is more beneficial.

5.2.2 Tuning as an Optimization Problem

The tuning problem can be easily formulated as a general optimization problem [42], where the objective function to *maximize* is the speech quality, or MOS, produced by

the SE system. Since most measures are full-referenced, we calculate the difference in MOS as

$$\Delta\text{MOS}\left(\hat{s}[n], y[n]\right) = \text{MOS}\left(\hat{s}[n], s[n]\right) - \text{MOS}\left(y[n], s[n]\right).$$

The SE system tuning can be formulated mathematically as a constrained optimization problem. Let $\hat{s}[n, \mathbf{p}]$ be the SE system output obtained with $\mathbf{p}$, the problem can be written as:

$$\begin{aligned} \text{maximize} \quad & \text{Q}(\hat{s}[n, \mathbf{p}]), \\ \text{subject to} \quad & \text{C}(\mathbf{p}) \leq C_{\max}, \end{aligned} \tag{180}$$

where $\text{Q}(\cdot)$ is the optimization criterion and $C_{\max}$ is the computational complexity constraint. Additionally, we can define $\mathbf{L}$ and $\mathbf{U}$ as the lower and upper bounds of $\mathbf{p}$, i.e., $\mathbf{L} \leq \mathbf{p} \leq \mathbf{U}$. Since the objective function is nonlinear and not known to be convex, there is no formal efficient solution (180). However, the nonlinear programming problem can still be practically solved by several approaches, each of which involves some compromises [10].

The genetic algorithms (GAs) have been successfully applied to this type of non-convex mixed-integer optimization problems [46]. The basic idea is to apply genetic operators, such as *mutation* and *crossover*, to *evolve* a set of initial solutions, or *population*, in order to find the solution that maximizes the objective function. The key element of this evolutionary process for dealing with the nonlinear constraints is the so-called *tournament selections*, which, unlike hill-climbing algorithms, allow for several random pairwise comparisons between sets of parameters and quickly determine the boundary of the feasible region [22]. The various steps of the algorithm are outlined below.

- **Step 1 -** An initial population of M solutions is first generated by randomly choosing the values of each set from the feasible region $\mathbf{p}_m^{(0)} \sim \mathcal{U}(\mathbf{L}, \mathbf{U})$. As a

general remark, the feasible region determined by the bounds in (180) is larger than the one allowed by the constraint, e.g., the complexity of the $\mathbf{U}$ solution might be much higher than $C_{\max}$. However, a methodology will be used in the evolutionary process to enforce the feasibility of the solution.

- **Step 2 -** The sets that go through crossover or mutation are chosen in a series of tournament selections: a random parameter set ω is extracted from the population, $\Omega \subset \mathbf{\Pi}^{(k)}$, and the set $\mathbf{p}_m^{(k)} \in \Omega$ with the best $\mathrm{Q}(\hat{s}[n, \mathbf{p}_m^{(k)}])$ is then selected. A constraint is imposed in the pairwise comparison of the tournament selection by making sure that when a feasible and an infeasible solutions are compared, the feasible one is chosen, and when two infeasible solutions are compared, the one with smaller constraint violation is chosen [22].

  Crossover - This operator allows to combine two sets of parameters with good but not optimum values of their objective function from a previous generation, $\mathbf{p}_n^{(k)}, \mathbf{p}_l^{(k)} \in \mathbf{\Pi}^{(k)}$, through a random weighted mean:

$$\mathbf{p}_m^{(k+1)} = \Phi(\mathbf{p}_n^{(k)}, \mathbf{p}_l^{(k)}) = \boldsymbol{\beta} \circ \mathbf{p}_n^{(k)} + (1 - \boldsymbol{\beta}) \circ \mathbf{p}_l^{(k)}, \tag{181}$$

  where $\boldsymbol{\beta} \sim \mathcal{U}(\mathbf{0}, \mathbf{1})$.

  Mutation - The mutation $\mathbf{p}_m^{(k+1)} = \Psi(\mathbf{p}_n^{(k)})$ of the set of values prevents choosing all elements in the population from a local minimum. Different heuristic approaches can be used, often associated with the type of the problem [24, 84]. The uniform perturbation is a simple operator that replaces the value of a l^{th} element with a uniform random value selected between the upper and lower bounds:

$$\Psi_a(p_{n_l}^{(k)}) = \delta, \quad \delta \sim \mathcal{U}\left(L_l, U_l\right). \tag{182}$$

- **Step 3 -** When a halting criterion is reached, the set of parameters that maximizes the objective function will be our solution:

$$\hat{\mathbf{p}} = \underset{\mathbf{p}_m^{(K)} \in \mathbf{\Pi}^{(K)}}{\arg\max}\, \mathrm{Q}\left(\hat{s}[n, \mathbf{p}_m^{(K)}]\right) \quad \text{s.t.} \quad \mathrm{C}(\mathbf{p}_m^{(K)}) \leq C_{\max}. \tag{183}$$

  Note that, given that not necessarily all the solutions in the K^{th} generation might fall within the feasible region [22], we choose the best solution that respects the constraint.

5.2.3 Database Generation

A key element to any data driven approach is to have a large and well structured amount of data for training and testing that correlates well to real world scenarios. The modeling of human conversational speech and the so-called conversational events, such as talk-spurt, pause, mutual silence, and double-talk, is fundamental to characterizing realistic scenarios in full-duplex communication. In particular, the studies done in [11] and [12] had a direct impact on the method for generating artificial conversational speech presented in [68]. However, this method is rather simplistic and relies on hand-coded expert knowledge [77], which is not easily transferable to the automatic generation of a large conversational speech database.

Several new methodologies have been proposed to model the turn-taking behaviors [96]. However, these methodologies are focused on human-machine turn-taking with very little mutual social interaction. We therefore focus on older studies on human-human conversations like [12]. In particular, we propose a flexible model of conversational behavior using a 4-state Markov chain model, where the states correspond to, respectively, mutual silence (MS), near-end (NE) talk, far-end (FE) talk, and double-talk (DT), and define all the possible combinations of the components in $y[n]$.

The Markov chain is uniquely described by its transition matrix $\mathbf{T}$ to model the generation model in [68] and the related distributions of the conversational events.

According to the distribution of the single talk duration, T_{ST}, the double talk duration, T_{DT}, and the mutual silence duration, T_{MS}, presented in [68], we are able to use a Markov chain Monte Carlo sampling algorithm [1] to find the transition matrix $\mathbf{T}$ of the 4-state Markov chain. Given that the transition between active NE and active FE and the transition between MS and DT are not allowed, and that the transition probabilities of going from MS to NE and MS to FE are equivalent [68], the Markov chain is uniquely represented by only four parameters:

$$\mathbf{T} = \begin{bmatrix} 1-2p_1 & p_1 & p_1 & 0 \\ p_2 & 1-p_3-p_2 & 0 & p_3 \\ p_2 & 0 & 1-p_3-p_2 & p_3 \\ 0 & p_4 & p_4 & 1-2p_4 \end{bmatrix}. \tag{184}$$

This makes it very easy to modify and fit different types of conversation scenarios with different levels of interactivity [50]. An example of a sequence of conversational speech and its Markov chain model is shown in Figure 36.

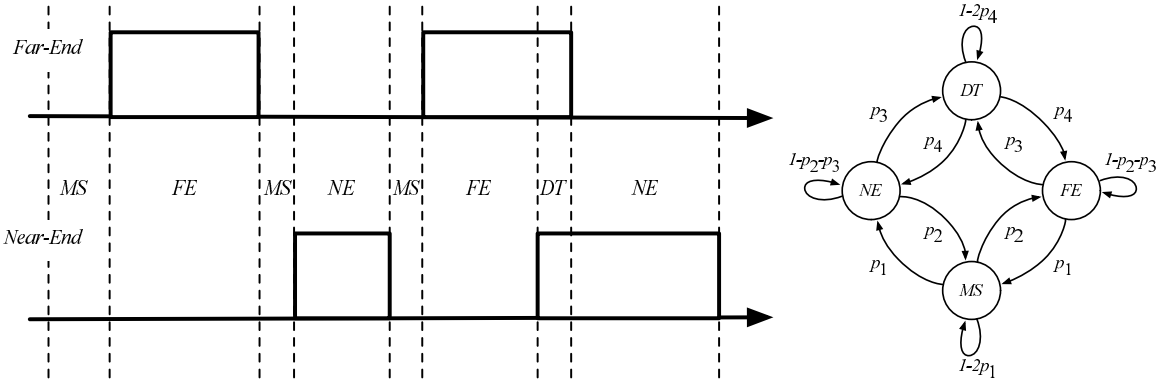


Figure 36: Conversational sequence and its Markov chain model.

5.2.4 Experimental Results

5.2.4.1 Unconstrained Optimization

The speech databases were generated using the ITU-T P-Series test signals [66]. This set includes 16 recorded sentences in each of 20 languages and sentences recorded in

an anechoic environment, sampled at 16 kHz. From these, we generated two single-channel signals, NE and FE, with continuous activity (i.e., without pauses). The total duration of the speech is about one hour per channel. The NE and FE speech segments were generated using the Markov chain with $p_1 = 0.04$, $p_2 = 0.03$, $p_3 = 0.05$, and $p_4 = 0.25$, generating the same statistical behavior of conversational events as specified in [68].

A noise database comprised of babble (e.g., airport, cafeteria, exhibition, and restaurant) noise, white and pink noise, impulsive noise (e.g., hammering), airplane cabin noise, car noise from a variety of car models, and street noise was used. The RIRs were calculated in office environments using the Audio Precision APx525 log-swept chirp signal through the *Beats Pill*™ portable speaker and truncated to the desired length (f_s= 48 kHz, resampled at 16 kHz). A set of 10 RIRs was then chosen with average reverberation time, RT_{60}, of 280 ms [101].

In order to generate the NE and FE segments, the starting and ending points were chosen randomly within the NE and FE channels. We generated 1000 segments with lengths between 6 to 8 s, ideal for objective quality measures [71, 72]. The two segments were then normalized to -26 dBov to avoid clipping, following the ITU-T Recommendation P.835 [70], and convolved with their respective RIR with normalized unitary energy. The microphone signal was created as follows. The NE signal was mixed with the FE signal at signal-to-echo ratios (SERs) uniformly distributed between -30 and 5 dB. The scaling was done by calculating the energy of the signals according to [67]. The noise was then mixed at an SNR uniformly distributed between -5 to 10 dB, according to the noise and the mixed speech signal energies [59].

The genetic algorithm had a population of $M = 20$ possible candidates, and the best $N = 4$ were migrated to the next generation. These values were chosen empirically to balance the complexity and the accuracy of the results. Of the remaining sets, half went through crossover and half went through mutation (uniform mutation

was chosen). The perceptual objective quality measure used was the average ΔMOS, as obtained through PESQ [71], POLQA [72], and the recently introduced Virtual Speech Quality Objective Listener (ViSQOL) [56, 57]. We included the manually tuned system, where the parameters were selected during the algorithmic design phase as a reference, and obtained four sets of parameters: $\mathbf{p}_{\text{POLQA}}$, $\mathbf{p}_{\text{PESQ}}$, and $\mathbf{p}_{\text{ViSQOL}}$, and $\mathbf{p}_{\text{MANUAL}}$. For comparison, we also optimized the SE system over four traditional objective measures, averaged over the evaluation set, that do not account for perception: log-spectral distortion (LSD), true echo return loss enhancement (TERLE), MSE, and a combined measure where the AEC block is optimized first using TERLE, and the RPE, NPE, and NS blocks are optimized with LSD (with fixed AEC parameters). The following sets were obtained with proposed optimization method: $\mathbf{p}_{\text{LSD}}$, $\mathbf{p}_{\text{TERLE}}$, $\mathbf{p}_{\text{MSE}}$, and $\mathbf{p}_{\text{TERLE+LSD}}$.

We divided the database into two parts, where 80% was used to estimate the parameters and 20% was used for testing. Table 24 shows the ΔMOS calculated using PESQ, POLQA, ViSQOL, and various traditional objective measures. The results show a net improvement in MOS over the manually tuned method, which in turn outperforms all the traditional objective measures. This proves that, in general, a trained ear is much better at determining proper values for the various parameters than using only the traditional objective measures, even if the tuning is done on a limited set. However, the use of perceptual objective measures for large-scale optimization greatly improves the performance of the SE algorithm over a much larger dataset. $\Delta\text{MOS}_{\text{POLQA}}$, arguably the most reliable measure for SE performance evaluation, shows that $\mathbf{p}_{\text{POLQA}}$ is .358 above $\mathbf{p}_{\text{MANUAL}}$ which is remarkable since there is no algorithmic modification other than using a better perceptual objective measure.

A subjective evaluation was performed through the MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA) test [64]. We compared the manually tuned

Table 24: Comparison between the objective improvements obtain with the SE algorithm in terms of MOS calculated with POLQA, PESQ, and ViSQOL obtained with different sets of parameters as result of optimizing with different criteria. A 95% confidence interval is given for each value.

method	$\Delta\text{MOS}_{\text{PESQ}}$	$\Delta\text{MOS}_{\text{POLQA}}$	$\Delta\text{MOS}_{\text{ViSQOL}}$
$\mathbf{p}_{\text{POLQA}}$	.455±.021	.654±.042	.387±.021
$\mathbf{p}_{\text{PESQ}}$	.475±.035	.442±.050	.342±.053
$\mathbf{p}_{\text{ViSQOL}}$	.358±.028	.487±.450	.369±.032
$\mathbf{p}_{\text{MANUAL}}$	.276±.083	.296±.121	.201±.089
$\mathbf{p}_{\text{LSD}}$	.139±.042	.221±.046	.154±.043
$\mathbf{p}_{\text{TERLE}}$	.147±.053	.234±.067	.121±.025
$\mathbf{p}_{\text{TERLE+LSD}}$	.194±.061	.246±.049	.173±.082
$\mathbf{p}_{\text{MSE}}$	.138±.089	.179±.134	.104±.091

configuration $\mathbf{p}_{\text{MANUAL}}$ with the two configurations obtained with standardized ITU-T tools, $\mathbf{p}_{\text{POLQA}}$ and $\mathbf{p}_{\text{PESQ}}$. The anchors were chosen as a 3.5 kHz low-pass filtered version (LP3.5) of the reference signal for scaling, as specified in [64], and the unprocessed speech, to represent the worst-case scenario in the listening evaluation. A pool of eleven expert listeners, familiar in detecting small impairments, and seven naive listeners was chosen. The test was performed using six speech excerpts randomly selected from the testing database. The results shown in Figure 37 are in line with the objective analysis. In particular, the confidence interval of the POLQA score only minimally overlaps with other scores, showing a significant statistical difference. The high variance of the LP3.5 and manually tuned scores is explained by the observed bimodality of the distribution of these scores, with a good percentage of the subjects preferring the bandlimitedness of the anchor over the manually tuned enhanced speech. Nonetheless, all subjects consistently preferred the POLQA-based tuning.

5.2.4.2 Constrained Optimization

A new database was generated for the constrained optimization and an instance of the database was created as follows. We concatenated two sentences, randomly

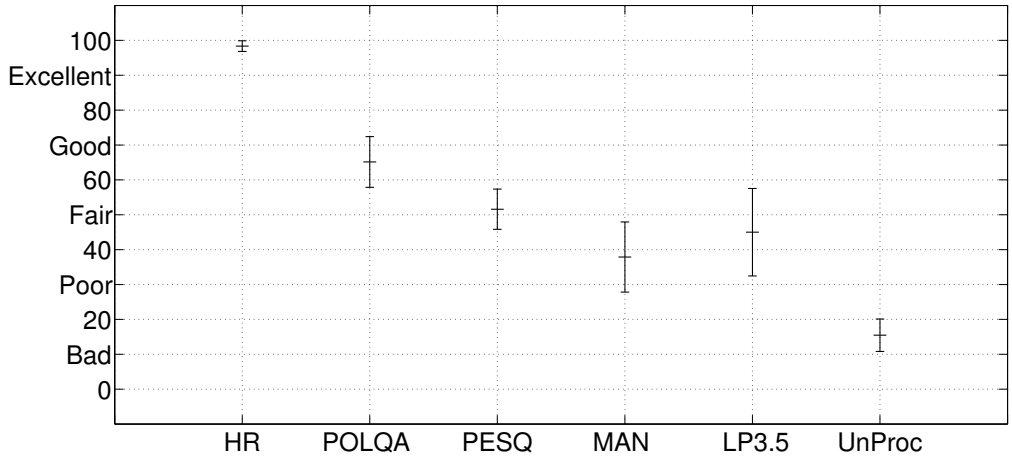


Figure 37: Results of the MUSHRA listening test comparing three different tuning strategies: manual tuning, PESQ-based, and POLQA-based.

chosen without replacement from a total of 6,300 TIMIT sentences to form the NE speech. We then extracted their voice activity from their phonetic transcription (given on a sample by sample basis) to determine the durations of the speech and non-speech parts. Since the TIMIT sentences have little non-speech sections, we randomly zero-padded the beginning and the end of the concatenated speech file as well as between the two TIMIT sentences so that the speech activity had a uniform duration distribution between 30% to 45% and the non-speech probability between 55% to 70%, in line with the studies on conversational speech presented in [11].

The FE speech pattern was generated using a 2-state Markov chain which is a collapsed version of the 4-state Markov chain used in [44], given that the NE pattern is already given. In particular, from the FE side, MS coincides with NE, creating a PAU state, and DT coincides with FE itself, creating a TS state. We tuned the transition probabilities in the transition matrix of the Markov chain to match the above mentioned statistics of the NE speech using a Markov chain Monte Carlo sampling algorithm [1]. The FE speech database was generated by concatenating and removing pauses from the ITU-T P-Series [66]. Once the on-off speech pattern of the FE was created, we randomly chose the starting and ending point in the FE

channel, and then we overlapped it with the NE. Given that certain transitions are not allowed in the conversational model [68], we ran several instances of the Markov chain until the DT probability ranges from 7% to 17%, the MS probability from 20% to 30%, and no DT-MS and NE-FE transitions occurred.

A noise database comprising of babble noise (e.g., airport, cafeteria, exhibition, and restaurant), white and pink noise, impulsive noise (e.g., hammering), airplane cabin noise, car noise from a variety of car models, and street noise was used. The RIRs were calculated in office environments using the Audio Precision APx525 log-swept chirp signal through the *Beats Pill*$^{\text{TM}}$ portable speaker and truncated to the desired length. A set of 10 RIRs was then chosen with an average reverberation time RT_{60} = 280 ms. The 3,150 NE and FE segments were then normalized to -26 dBov to avoid clipping by following the ITU-T Recommendation P.835 [70], and convolved with their respective RIR with normalized unitary energy. The NE signal was mixed with the FE signal at signal-to-error ratio (SER) uniformly distributed between -30 and 5 dB. The scaling was done by calculating the energy of the signals according to [67]. The noise was then mixed at an SNR uniformly distributed between -5 to 10 dB, according to the noise and the mixed speech signal energies [59]. The choices of RIRs, SER, and SNR were considered empirically appropriate given the possible usage scenarios for a portable teleconferencing device.

For the ASR front-end scenario, the capability of the recognizer were examined by measuring its accuracy in recognizing phones, the building blocks of words and utterances [94], through PAR. We used the HTK toolkit [130] to train an acoustic model composed of 61 phones [38]. A set of 13 mel-frequency cepstral coefficients (MFCCs) with their first and second derivatives, for a total of 39 coefficients, were generated and used as features for our experimental analysis. We used a 5-state HMM with an 8-mixture Gaussian mixture model (GMM) for each phone, a fairly standard setup [94]. We normalized the mean of the MFCCs as suggested in [52] for the proper

application of the direct masking. We trained our HMMs with clean speech only to focus only on the SE capabilities.

For the optimization problem in (180), the total complexity was fixed to $C_{\max} = 50$ MCPS. The genetic algorithm had a population of $M = 100$ possible candidates and $K = 10$ generations, which we observed to be a good trade-off between the accuracy of the solution and the duration of the optimization process. Given the relatively small size of the population, we chose a deterministic tournament selection [15] by calculating the fitness function $\mathrm{Q}(\,\cdot\,)$ for all the elements of the population. A seed was given to generate the initial population by biasing this towards a known hand-tuned solution that achieved reasonable values in the algorithmic design phase, $\mathbf{p}_{\mathrm{INIT}}$. This was done with the same operator used in the crossover operation (181), where each randomly generate solution is weighted with $\mathbf{p}_{\mathrm{INIT}}$ and $\boldsymbol{\beta} \sim \mathcal{U}(\mathbf{0.3}, \mathbf{0.7})$. The best $M = 20$ or less sets of parameters in each generation that fulfill the constraint were migrated to the next generation, of the remaining sets half went through crossover and half through mutation. The optimization process took about 90 hours on a 16-core Intel Xeon machine with parallelized scripts. Note that while the tuning database is fixed, calculating $\mathrm{Q}(\hat{s}[n, \mathbf{p}])$ requires running all 3,150 signals for each population element $\mathbf{p}$ at each iteration. The analysis-modification-synthesis as well as the different algorithmic components operated on a 16 ms frame size (256 samples at 16 kHz) with 50% overlap.

The scatterplots of the fitness values $\mathrm{Q}(\,\cdot\,)$ for each element of the initial population and final population of the evolutionary optimization process are shown in Figure 38. The solution optimized for PAR, $\hat{\mathbf{p}}_{\mathrm{PAR}}$, and the solution optimized for MOS, $\hat{\mathbf{p}}_{\mathrm{MOS}}$, on the training database not only achieve much higher PAR and ΔMOS but also achieve a net 20% reduction in computational complexity. The unconstrained solutions are also calculated, $\hat{\mathbf{p}}_{\mathrm{PAR_u}}$ and $\hat{\mathbf{p}}_{\mathrm{MOS_u}}$, respectively. The final sets of parameters are chosen according to (183) and evaluated on the testing database. The results

are shown in Table 25.

Table 25: Results of the GA optimization algorithm on the testing database.

	PAR [%]	ΔMOS	C (**p**) [MCPS]
$\mathbf{p}_{\text{INIT}}$	51.04	0.32	49.14
$\hat{\mathbf{p}}_{\text{PAR}}$	62.94	0.65	41.17
$\hat{\mathbf{p}}_{\text{PAR}_{\text{u}}}$	63.15	0.68	53.56
$\hat{\mathbf{p}}_{\text{MOS}}$	60.07	0.87	42.56
$\hat{\mathbf{p}}_{\text{MOS}_{\text{u}}}$	60.22	0.92	55.23

While similar mean fitness values for the the last population $\mathbf{\Pi}^{(K)}$ and its immediate preceding ones, i.e., $\mathbf{\Pi}^{(K-1)}$ proved overall convergence, it was observed the existence of quasi-optimal solutions within the final population that can have significantly different element-wise values, $p_{ml}^{(K)}$. This *non-uniqueness* problem, often encountered in nonlinear programming, [10] is, arguably, not a weakness in our case. In fact, having a set of possible candidates with different characteristics increases our chances of determining a set of parameters that offers better properties for our purposes, while still moving in the neighborhood of the optimal value. In this regard, we have observed that the seed for the generation of the initial population given by $\hat{\mathbf{p}}_{\text{INIT}}$, did not affect the values of the final fitness function or the overall behavior of the final population. In fact, $\hat{\mathbf{p}}_{\text{INIT}}$ biased the initial population but did not restrict the actual search region determined by the bounds $\mathbf{L}$ and $\mathbf{U}$. The major impact of pointing the search towards a reasonable path results in speeding up the genetic algorithm convergence by reducing both number of iterations and size of the population.

In informal listening, the difference in the output processed with $\hat{\mathbf{p}}_{\text{PAR}}$ and $\hat{\mathbf{p}}_{\text{MOS}}$, follow known differences in SE when targeting recognition and intelligibility versus perceived quality of speech. A clear example is the binary mask being enabled by the optimization process only in the $\hat{\mathbf{p}}_{\text{PAR}}$, while the $\hat{\mathbf{p}}_{\text{MOS}}$ solution exploited the perceptual masking properties of speech in noisy conditions.

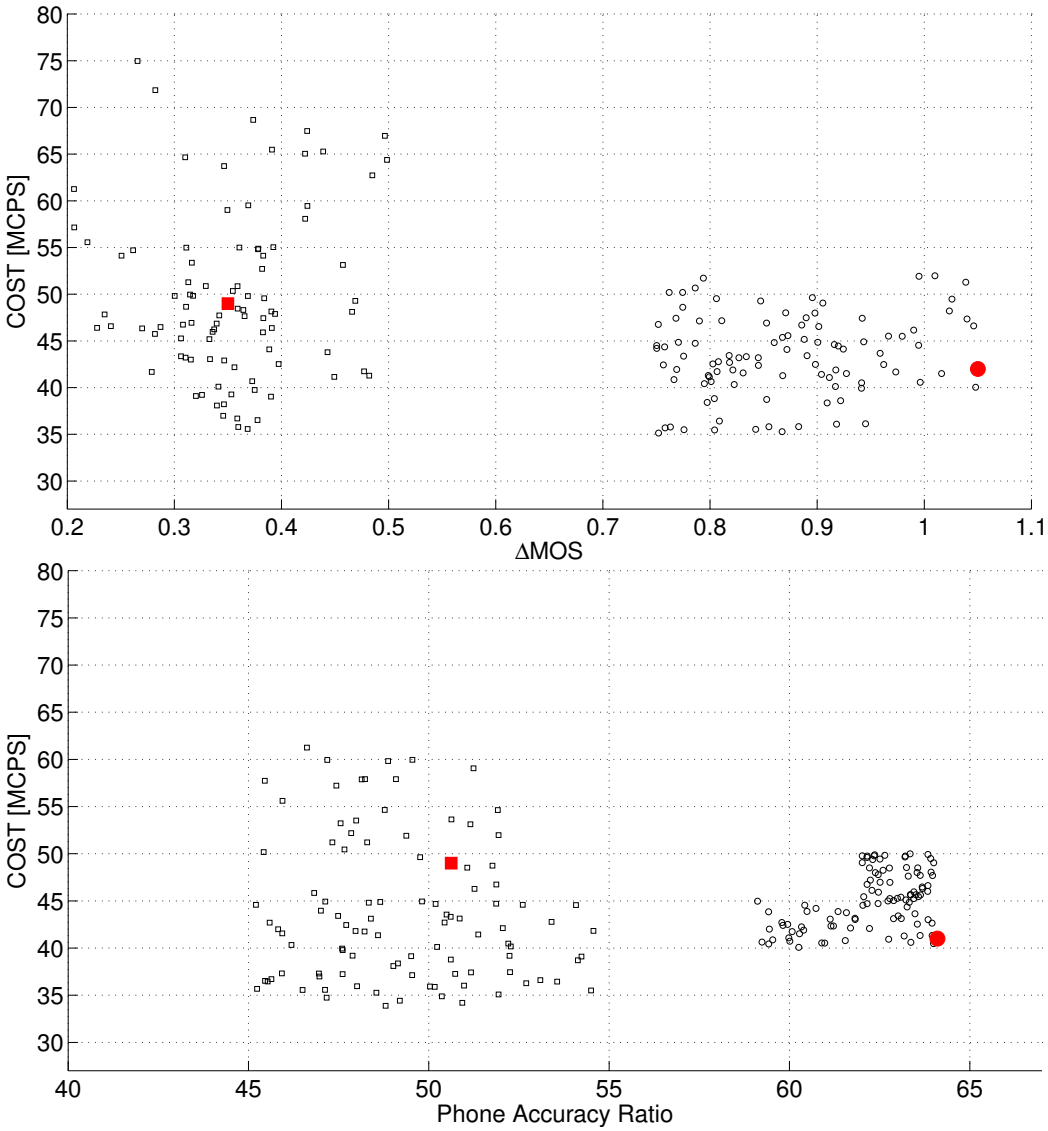


Figure 38: Results of the GA on the training database. Initial population (squares) and final population (circles) in the constrained optimization over ΔMOS and PAR on the training database. The initial solution $\mathbf{p}_{\text{INIT}}$ is the red square, while the optimal final solution that respects the constraint is the red circle.

5.3 Robust Stereophonic Echo Canceler

5.3.1 Robust Regularization for Stereophonic Adaptive Filter

The two-channel FDAF [5, 34, 75] used in Section 4.4.1 was shown to achieve fast convergence with the proposed decorrelation by sub-band resampling introduced in Chapter 4. The fixed regularization term in Table 15, however, assumed that the near-end interference, i.e., the echo-to-noise ratio (ENR), was known *a priori* and fixed at a constant level. In practical scenarios, the near-end interference, or near-end speech, is constantly varying across time and frequency. Therefore, using a fixed regularization term is no longer adequate for a robust performance of the AEC under nonstationary near-end interference. Here we propose a modified two-channel frequency-domain adaptive filter (FDAF) to adjust the regularization term according to the level of near-end interference.

Recall that the regularized power spectral densitys (PSDs) in Table 15 is equal to

$$\tilde{\mathbf{S}}_{pp}[m] = \hat{\mathbf{S}}_{pp}[m] + \delta \mathbf{I}_{2L\times 2L}, \quad p = 1, 2, \tag{185}$$

where L is the adaptive filter length and δ is the regularization term that is equal to

$$\delta = \frac{L(1+\sqrt{1+\mathrm{ENR}})}{\mathrm{ENR}}(\sigma_{x_1}^2 + \sigma_{x_2}^2). \tag{186}$$

For loudspeaker-enclosure-microphon (LEM) devices, e.g., *Beats Pill*$^{\mathrm{TM}}$, the ENR is typically very high at around 25 to 35 dB due to the close proximity of the loudspeakers to the microphones. Therefore, the frequency domain regularization term in (186) can be approximated as

$$\begin{aligned}\delta(\omega) &\approx \gamma \frac{S_{x_1x_1}(\omega) + S_{x_2x_2}(\omega)}{\sqrt{\mathrm{ENR}}} \\ &= \gamma(S_{x_1x_1}(\omega) + S_{x_2x_2}(\omega))\sqrt{\frac{S_{ss}(\omega) + S_{vv}(\omega)}{S_{dd}(\omega)}},\end{aligned} \tag{187}$$

where γ is a constant factor and S_{ss}, S_{vv}, and S_{dd} are the PSDs of the near-end speech, near-end noise, and the echo, respectively. The last equality in (187) assumes

the near-end speech and noise are uncorrelated.

The PSD of the near-end echo can be rewritten as (omitting the frequency index)

$$\begin{aligned} S_{dd} &= \mathrm{E}\{(D_1 + D_2)(D_1 + D_2)^*\} \\ &= |H_1|^2 S_{x_1x_1} + |H_2|^2 S_{x_2x_2} + \mathrm{E}\{H_1 H_2^* X_1 X_2^*\} + \mathrm{E}\{H_2 H_1^* X_2 X_1^*\} \\ &\approx |H_1|^2 S_{x_1x_1} + |H_2|^2 S_{x_2x_2} \end{aligned} \tag{188}$$

where the last equality assumes that the reference signals are sufficiently decorrelated, i.e., $\mathrm{E}\{X_i X_j^*\} \approx 0, i \neq j$. If we further assumes that the two loudspeakers and the microphones are closely and symmetrically spaced such that $H_1 \approx H_2 \approx H$, (187) can be further simplified as (again omitting ω)

$$\begin{aligned} \delta &\approx \gamma \frac{\sqrt{(S_{x_1x_1} + S_{x_2x_2})(S_{ss} + S_{vv})}}{\sqrt{2|H|^2}} \\ &\approx \gamma' \sqrt{(S_{x_1x_1} + S_{x_2x_2}) S_{ee}}, \end{aligned} \tag{189}$$

where the last approximation assumes that the RIRs vary slowly compared to the PSDs such that the frequency-domain RIRs can be approximated as a constant term and be absorbed by the constant γ. Although constant term γ in theory should be a frequency dependent term that approximates the statics of the RIRs, here we assume no prior knowledge of the RIRs and choose a fixed γ across all frequency bins. We also assumes that the adaptive filter is sufficiently converged such that the error PSD is approximately equal to the PSD of the near-end speech plus noise. The final two-channel FDAF algorithm with robust regularization is summarized in Table 26.

The simulation setup was the same as Section 4.4.1 for direct comparison, and all parameter values were also set to be the same as before except for the robust regularization and $\gamma = 1$. Refer to Section 4.4.1 for the detailed description of the parameter values. The decorrelation procedures and the parameters were the same as Section 4.4.3. Figure 39 shows the misalignment performance using the proposed robust regularization term, where the near-end interference is white Gaussian noise (WGN)

Table 26: The two-channel FDAF with robust regularization.

Definitions

$$
\begin{aligned}
[\mathbf{F}]_{k+1,n+1} &= e^{-j\frac{\pi}{L}kn}, \quad k,n = 0,\ldots,2L-1 \\
\mathbf{G}^{01} &= \mathbf{F}\begin{bmatrix}\mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{L\times L} & \mathbf{0}_{L\times L}\\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L}\end{bmatrix}\mathbf{F}^{-1} \\
\underline{\mathbf{w}}_p &= \mathbf{F}[\mathbf{w}_p^{\mathrm{T}}, \mathbf{0}_{L\times 1}^{\mathrm{T}}]^{\mathrm{T}}, \quad p = 1,2 \\
\mu' &= \mu(1-\lambda),\ 0 < \mu \leq 1,\ 0 \ll \lambda < 1, \gamma > 0
\end{aligned}
$$

Spectral estimation

$$
\begin{aligned}
\mathbf{X}_p[m] &= \mathrm{diag}\{\mathbf{F}[x_p[(m-1)L],\ldots,x_p[(m+1)L-1]]^{\mathrm{T}}\}, \quad p = 1,2 \\
\underline{\mathbf{y}}[m] &= \mathbf{F}[\mathbf{0}_{1\times L}, y[mL],\ldots,y[(m+1)L-1]]^{\mathrm{T}} \\
\underline{\mathbf{e}}[m] &= \underline{\mathbf{y}}[m] - \mathbf{G}^{01}(\mathbf{X}_1[m]\mathbf{w}_1[m-1] + \mathbf{X}_2[m]\mathbf{w}_2[m-1]) \\
\mathbf{E}[m] &= \mathrm{diag}\{\underline{\mathbf{e}}[m]\} \\
\hat{\mathbf{S}}_{ee}[m] &= \lambda\hat{\mathbf{S}}_{ee}[m-1] + (1-\lambda)\mathbf{E}[m]\mathbf{E}^*[m] \\
\hat{\mathbf{S}}_{ij}[m] &= \lambda\hat{\mathbf{S}}_{ij}[m-1] + (1-\lambda)\mathbf{X}_i[m]\mathbf{X}_j^*[m], \quad i,j = 1,2 \\
\boldsymbol{\Delta}[m] &= \gamma[(\hat{\mathbf{S}}_{11}[m] + \hat{\mathbf{S}}_{22}[m])\hat{\mathbf{S}}_{ee}[m]]^{\circ\frac{1}{2}} \\
\tilde{\mathbf{S}}_{pp}[m] &= \hat{\mathbf{S}}_{pp}[m] + \boldsymbol{\Delta}[m], \quad p = 1,2 \\
\hat{\mathbf{C}}_{12}[m] &= (\tilde{\mathbf{S}}_{11}[m]\tilde{\mathbf{S}}_{22}[m])^{-1}\hat{\mathbf{S}}_{12}[m]\hat{\mathbf{S}}_{21}[m] \\
\hat{\mathbf{S}}_p[m] &= \tilde{\mathbf{S}}_{pp}[m](\mathbf{I}_{2L\times 2L} - \hat{\mathbf{C}}_{12}[m]), \quad p = 1,2 \\
\mathbf{K}_1[m] &= \hat{\mathbf{S}}_1^{-1}[m](\mathbf{X}_1[m] - \hat{\mathbf{S}}_{12}[m]\tilde{\mathbf{S}}_{22}^{-1}[m]\mathbf{X}_2[m]) \\
\mathbf{K}_2[m] &= \hat{\mathbf{S}}_2^{-1}[m](\mathbf{X}_2[m] - \hat{\mathbf{S}}_{21}[m]\tilde{\mathbf{S}}_{11}^{-1}[m]\mathbf{X}_1[m])
\end{aligned}
$$

Filter adaptation

$$
\underline{\mathbf{w}}_p[m] = \underline{\mathbf{w}}_p[m-1] + 2\mu'\mathbf{G}^{10}\mathbf{K}_p^*[m]\underline{\mathbf{e}}[m], \quad p = 1,2
$$

at ENR = 30 dB. Compared to Figure 33, we notice a similar convergence behavior between additive white Gaussian noise (AWGN), nonlinear processor (NLP), phase modulation (PMod), and sub-band resampling (SBR), and the steady-state misalignment is around −25 dB. The proposed SBR still achieves the fastest convergence rate.

Figures 40 and 41 show the misalignment performance comparison using a fixed δ

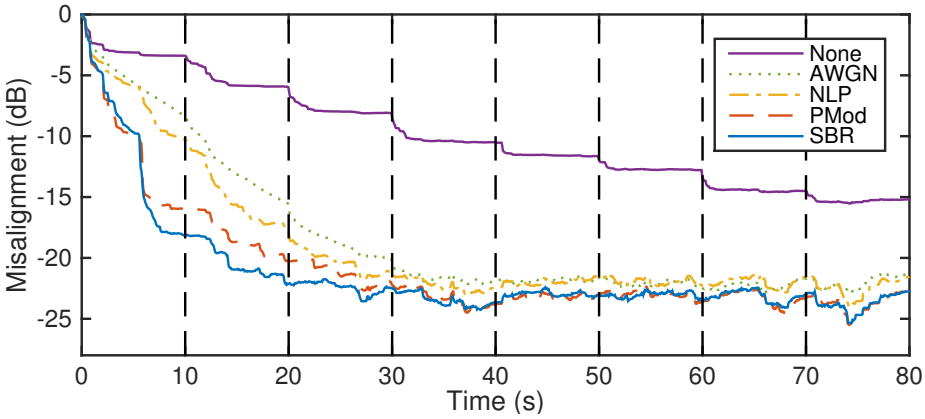


Figure 39: Misalignment comparison with robust regularization according to Table 26. The near-end interference is WGN at an ENR = 30 dB. The vertical dotted lines represent the instances when the far-end source location is changed.

and the proposed robust regularization, respectively, when the near-end interference is a speech signal. The near-end speech signals were mixed at ENR = 30 dB on average. We note that when using a fixed δ, the FDAF quickly diverges, especially when the far-end echo path or the near-end instantaneous ENR changes. The proposed SBR still reconverges much faster than other decorrelation methods each time the FDAF diverges. On the other hand, the proposed robust regularization achieves a much more stable misalignment performance even as the near-end speech level changes.

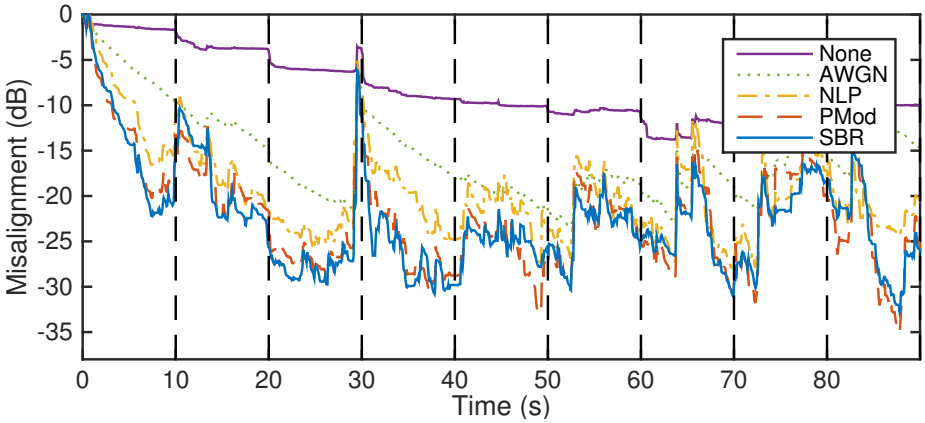


Figure 40: Misalignment comparison with fixed δ according to Table 15. The near-end interference is speech at an averaged ENR = 30 dB. The vertical dotted lines represent the instances when the far-end source location is changed.

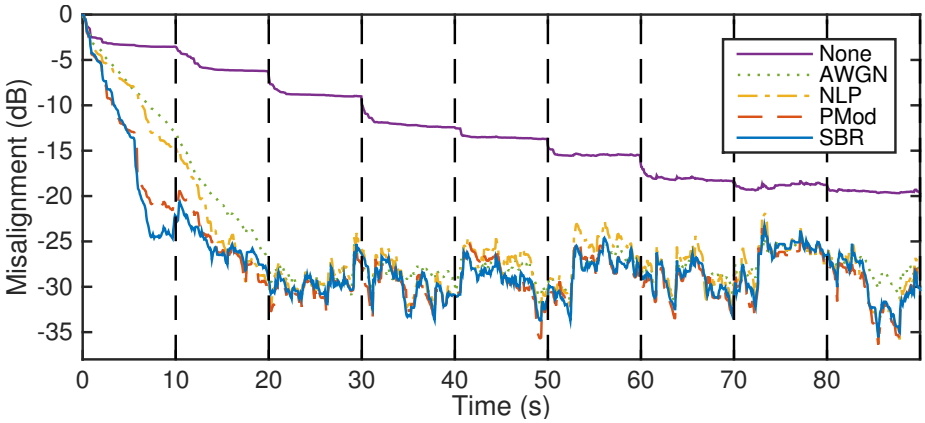


Figure 41: Misalignment comparison with robust regularization according to Table 26. The near-end interference is speech at an averaged ENR = 30 dB. The vertical dotted lines represent the instances when the far-end source location is changed.

Figure 42 shows the same graph as Figure 41 but zoomed out to show the steady-state convergence behavior when no decorrelation is applied. We notice that the steady-state misalignment is around -25 dB, which is the same as the theoretical prediction in Figure 33. The large fluctuation of the misalignment is likely due to the fact the the near-end ENR fluctuates greatly such that the FDAF converges to a smaller error when the instantaneous ENR is higher than the nominal value of 30 dB. Still, with the proposed robust regularization and the proposed SBR, the steady-state misalignment rarely goes above -25 dB and is almost always better than other decorrelation methods in terms of the convergence and the misalignment performance.

5.3.2 Robust Stereophonic Multi-Delay Adaptive Filter

The stereophonic AEC discussed in the previous section is still the block-based algorithm, which introduces large algorithmic delay. Therefore, we introduce the MDF version of the stereophonic AEC with robust regularization. To make the algorithm even more robust to the near-end interference, we also introduce the ERN and iterative adaptation that was used in the previous sections. Again due to the computational complexity constraint of the target platform, we elect to use the alternative

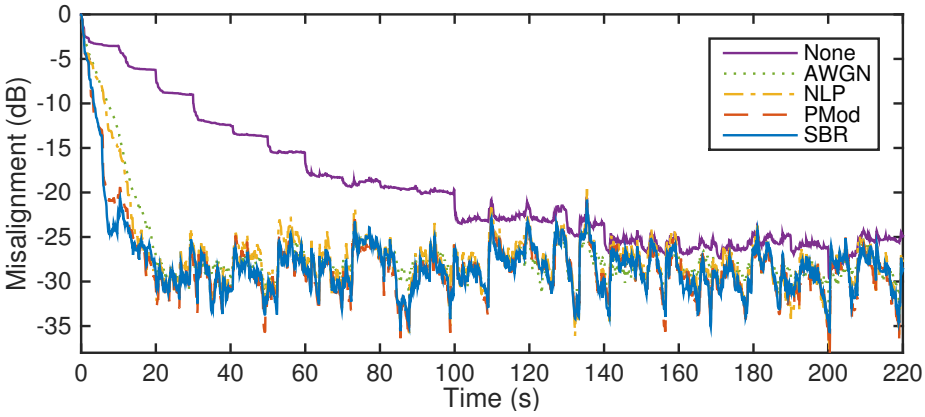


Figure 42: Misalignment comparison with robust regularization according to Table 26. Same configuration as Figure 41 but zoomed out to show the steady-state behavior.

constrained adaptation so that the gradient constraint is only applied to one partition per iteration. Table 27 summarizes the two-channel robust MDF algorithm with robust regularization, ERN, and iterative adaptation.

The simulation setup was again the same as Section 5.3.1 for direct comparison with all parameter values were also set to be the same as before except for $\mu = 0.25$, the frame-shift size $N = 256$, the number of partitions $M = 4$, and the number of iterations `numIterations` $= 1$. The decorrelation procedures and the parameters were again the same as Section 5.3.1. Figure 43 shows the misalignment performance using the MDF with the proposed robust regularization term, where the near-end interference is WGN at ENR = 30 dB. Compared to Figure 39, we notice a slightly slower convergence rate. This is due to the alternate constrained adaptation and the ERN that inevitably slow down the convergence rate. However, the alternate constrained adaptation is required for computational complexity reasons. Overall, the convergence behavior between different decorrelation procedures remains similar as Section 5.3.1.

Figures 44 and 45 show the misalignment performance comparison using a fixed δ and the proposed robust regularization with MDF, respectively, where the near-end interference is a speech signal. The near-end speech signals were mixed at an averaged

Table 27: The two-channel robust MDF algorithm.

Definitions

$$[\mathbf{F}]_{k+1,n+1} = \exp(-j\tfrac{\pi kn}{N}), \quad k,n = 0,\ldots,2N-1$$

$$\mathbf{G}^{01} = \mathbf{F}\begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}, \quad \mathbf{G}^{10} = \mathbf{F}\begin{bmatrix}\mathbf{I}_{N\times N} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N}\end{bmatrix}\mathbf{F}^{-1}$$

Initialization

$$\{\underline{\mathbf{w}}_{p,l} = \mathbf{0}_{2N\times 1},\, p = 1,2,\, l = 0,\ldots,M-1\}, \quad \{x[n] = 0;\, n < 0\}$$
$$\{\underline{\mathbf{s}}_{ij} = \mathbf{0}_{2N\times 1},\, i,j = 1,2\}, \quad \underline{\mathbf{s}}_{ee} = \mathbf{0}_{2N\times 1}$$
$$MN = L, \quad N > 0, \quad 0 < \mu \le 1, \quad \gamma > 0, \quad 0 < \epsilon \ll 1, \quad 0 \ll \beta < 1, \quad \mu' = \mu(1-\beta)$$

Iterative filter adaptation

$$\underline{\mathbf{X}}_p[m] = \mathrm{diag}\{\mathbf{F}[x_p[(m-1)N],\ldots,x_p[(m+1)N-1]]^{\mathrm{T}}\}, \quad p = 1,2$$
$$\underline{\mathbf{y}}[m] = \mathbf{F}[\mathbf{0}_{1\times L}, y[mN],\ldots,y[(m+1)N-1]]^{\mathrm{T}}$$

`for` $i :=$ `1 to numIterations`

$$\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \mathbf{G}^{01}\left(\sum_{p=1}^{2}\sum_{l=0}^{M-1}\underline{\mathbf{X}}_p[m-l]\underline{\mathbf{w}}_{p,l}\right)$$
$$\underline{\mathbf{E}}[m] = \mathrm{diag}\{\underline{\mathbf{e}}[m]\}$$
$$\underline{\mathbf{S}}_{ij} \leftarrow \beta\underline{\mathbf{S}}_{ij} + (1-\beta)(\underline{\mathbf{X}}_i[m]\underline{\mathbf{X}}_j^*[m]), \quad i,j = 1,2$$
$$\underline{\mathbf{S}}_{ee} \leftarrow \beta\underline{\mathbf{S}}_{ee} + (1-\beta)(\underline{\mathbf{E}}[m]\underline{\mathbf{E}}^*[m])$$
$$\underline{\boldsymbol{\Delta}} = \gamma[(\underline{\mathbf{S}}_{11}+\underline{\mathbf{S}}_{22})\underline{\mathbf{S}}_{ee}]^{\circ\frac{1}{2}}$$
$$\tilde{\underline{\mathbf{S}}}_{pp} = \underline{\mathbf{S}}_{pp} + \underline{\boldsymbol{\Delta}}, \quad p = 1,2$$
$$\underline{\mathbf{C}}_{12} = (\tilde{\underline{\mathbf{S}}}_{11}\tilde{\underline{\mathbf{S}}}_{22})^{-1}\underline{\mathbf{S}}_{12}\underline{\mathbf{S}}_{21}$$
$$\underline{\mathbf{S}}_p = \tilde{\underline{\mathbf{S}}}_{pp}(\mathbf{I}_{2N\times 2N} - \underline{\mathbf{C}}_{12}), \quad p = 1,2$$
$$\underline{\mathbf{K}}_{1,l} = \underline{\mathbf{S}}_1^{-1}(\underline{\mathbf{X}}_1[m-l] - \underline{\mathbf{S}}_{12}\tilde{\underline{\mathbf{S}}}_{22}^{-1}\underline{\mathbf{X}}_2[m-l]), \quad l = 0,\ldots,M-1$$
$$\underline{\mathbf{K}}_{2,l} = \underline{\mathbf{S}}_2^{-1}(\underline{\mathbf{X}}_2[m-l] - \underline{\mathbf{S}}_{21}\tilde{\underline{\mathbf{S}}}_{11}^{-1}\underline{\mathbf{X}}_1[m-l]), \quad l = 0,\ldots,M-1$$
$$\phi(E_k[m]) = \begin{cases}\sqrt{S_{ee}[k]}\, e^{j\angle E_k[m]}, & |E_k[m]| > \sqrt{S_{ee}[k]}\\ E_k[m], & \text{otherwise}\end{cases}$$
$$\underline{\mathbf{w}}_{p,l} \leftarrow \begin{cases}\underline{\mathbf{w}}_{p,l} + 2\mu'\mathbf{G}^{10}\underline{\mathbf{K}}_{p,l}^*\phi(\underline{\mathbf{e}}[m]), & m \bmod M = l\\ \underline{\mathbf{w}}_{p,l} + 2\mu'\underline{\mathbf{K}}_{p,l}^*\phi(\underline{\mathbf{e}}[m]), & \text{otherwise}\end{cases}, \quad p = 1,2$$

`end for`

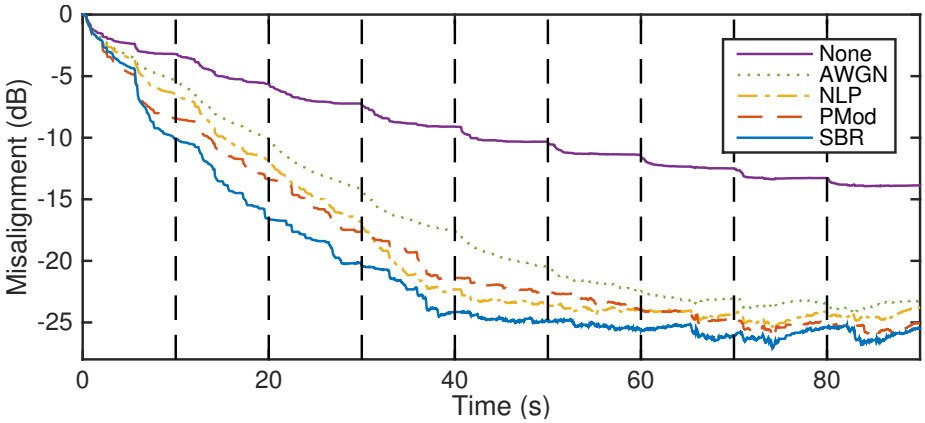


Figure 43: Misalignment comparison with robust regularization according to Table 27. The near-end interference is WGN at an ENR = 30 dB. The vertical dotted lines represent the instances when the far-end source location is changed.

ENR = 30 dB. We note again in Figure 44 the adaptive filter quickly diverges, and the divergence is worst with PMod, in some instances even worse then no decorrelation. Therefore, the fixed regularization term δ is unsuitable for real-world scenarios where the near-end interference is speech and the ENR constantly and drastically changes. The proposed robust regularization, on the other hand, provides stable adaptation as shown in Figure 45, and the proposed SBR still achieves the fastest convergence rate.

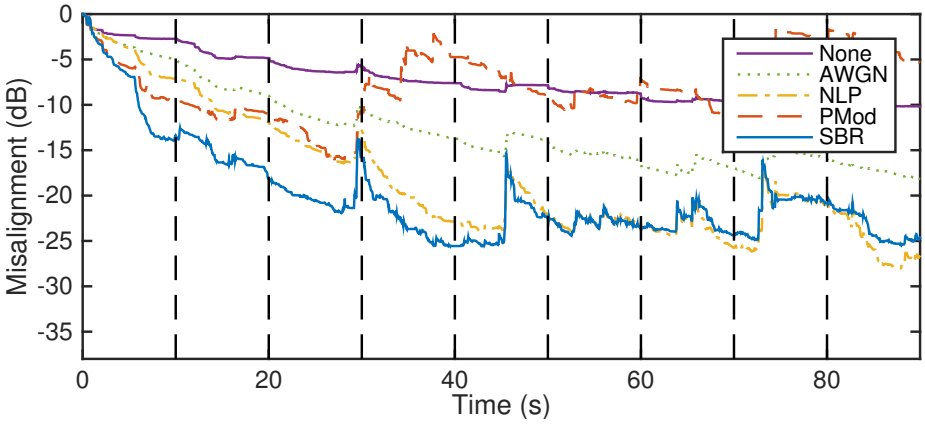


Figure 44: Misalignment comparison with fixed δ according to Table 15. The near-end interference is speech at an averaged ENR = 30 dB. The vertical dotted lines represent the instances when the far-end source location is changed.

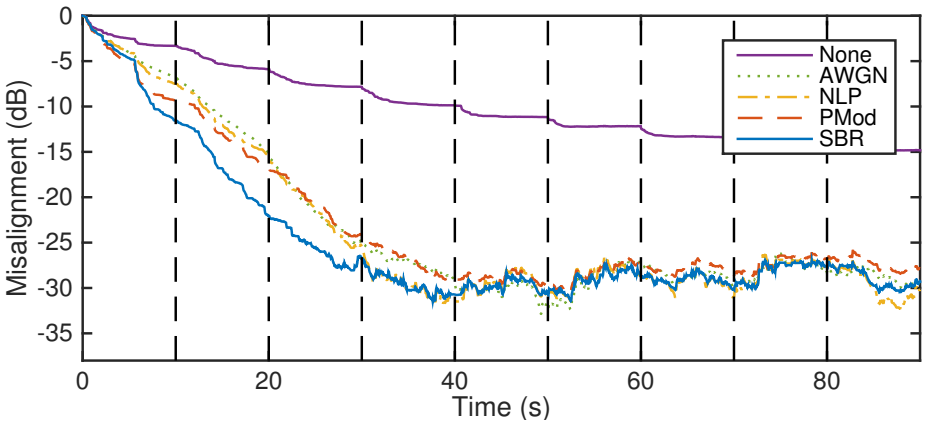


Figure 45: Misalignment comparison with robust regularization according to Table 27. The near-end interference is speech at an averaged ENR = 30 dB. The vertical dotted lines represent the instances when the far-end source location is changed.

CHAPTER VI

CONCLUSION

We conclude by reviewing the main ideas presented in this dissertation.

First, we presented several methods to systematically combine the acoustic echo cancellation (AEC) and residual echo suppression (RES) to improve the echo reduction results. Traditionally, the AEC and the RES are developed with different criteria and are treated as two separate units. The two units normally operates in their own framework that has no commonality or shared information between them. Through the system approach, the two units are combined in such a way that the performance of one of the unit can be further enhanced given the processed signal from the other unit. We discussed the system approach to RES, where the RES unit can be enhanced by using the knowledge of the linearly estimated echo signal from the AEC. We took the other direction and utilized the enhanced speech signal from the RES unit to boost the performance of the AEC. We showed how the system approach works in the *Kinect*™ audio system to demonstrate the benefit of the system approach.

Next, we proposed a perceptually motivated decorrelation procedure based on the system approach to alleviate the non-uniqueness problem during multi-channel acoustic echo cancellation (MCAEC) due to the correlation between reference signals that degrades the convergence performance of adaptive filtering algorithms. Although applying a decorrelation procedure before near-end playback can improve the tracking of the signal responses along the echo paths, the audio quality is usually compromised with the traditional decorrelation procedures. Furthermore, these decorrelation techniques may not achieve an optimal steady-state performance of an adaptive filter and

are usually performed in a full-band manner, leaving no possibility for frequency-selective decorrelation based on human perception. The resultant audio quality is traditionally examined in an ex post facto manner rather than an active factor in the design of the decorrelation procedure.

We proposed a sub-band resampling (SBR) decorrelation procedure motivated by the analysis of the sampling rate mismatch problem inherent to audio processing using distributed audio devices. The proposed SBR enables finer control and trade-off between the convergence rate of an adaptive filter and the audio quality, where the amount of resampling can be arbitrarily controlled in each frequency bin such that the perceptually less significant sub-bands can be more aggressively decorrelated while still preserving a high speech quality. We presented a thorough analysis of the inter-channel decorrelation by resampling for MCAEC. By smoothly varying the resampling ratio per frequency bin to achieve the desired coherence in each sub-band in order to minimize the signal distortion of the low frequency signal components, the proposed SBR scheme delivers consistently higher signal quality after the processing than other conventional decorrelation methods. Besides the superior speech quality, as measured by the high objective quality score and verified by a subjective listening test, a fast convergence rate is achieved with SBR.

Finally, we proposed an automated tuning framework for optimizing the whole voice system in a wide variety of acoustic mixing environments with different levels of signal-to-noise ratios (SNRs) and under various complexity considerations. Traditional approach to system tuning is usually done by expert tuners using a small set of database. However, this approach quickly becomes inadequate as the system complexity increases with many different components and tuning parameters. An alternative automated database generation and tuning approach was presented and discussed. Furthermore, due to the computational complexity constraint of a target platform, a constrained optimization framework was presented such that the tuned

parameter set maximizes the voice quality and is still feasible on the target platform. We verified experimentally the proposed tuning framework on the *Beats Pill*$^{\text{TM}}$ and demonstrated results of the tuning on either the objective voice quality or the automatic speech recognition (ASR) performance. Last but not least, we presented a robust stereo echo canceler using the proposed SBR decorrelation procedure and proposed an improved robust regularization procedure for the stereo echo canceler.

We list below our specific contributions from this dissertation.

- Proposed a system approach to RES and AEC and demonstrated the benefit of the system approach on the *Kinect*$^{\text{TM}}$ audio pipeline.

- Analyzed the decorrelation by resampling technique and showed the link between the resampling and the level of decorrelation measured in terms of coherence.

- Derived new closed-form expressions to demonstrate how resampling affects the steady-state misalignment performance of a stereophonic acoustic echo cancellation (SAEC).

- Proposed the perceptually motivated SBR technique, where the amount of decorrelation can be finely controlled for each sub-band, to alleviate the non-uniqueness problem with minimal distortion to the audio quality.

- Proposed an automated tuning procedure for a system-based approach towards echo cancellation performance and objective voice quality.

- Formulated the tuning prodecure as a constrained optimization problem to maximize the voice quality under a target computational complexity constraint.

- Proposed a robust regularization procedure for the SAEC that is low-delay and low complexity, suitable for implementation on embedded systems.

APPENDIX A

DISCRETE FOURIER TRANSFORM COEFFICIENTS OF WHITE GAUSSIAN NOISE AFTER RESAMPLING: WITHOUT FAR-END ROOM IMPULSE RESPONSE

Let $u[n]$ be a zero-mean white Gaussian noise (WGN) with variance σ_u^2. To fit the overlap-save frequency-domain adaptive filter (FDAF) structure in [102], here we assume the discrete short-time Fourier transform (STFT) of $u[n]$ is calculated every $2L$ samples with a frame shift of L, i.e.,

$$U_m[l] = e^{-j\pi lm} \sum_{n=0}^{2L-1} u_m[n] e^{-j\frac{\pi}{L}ln}, \tag{190}$$

where $l = -L, \ldots, L-1$, $u_m[n] = u[n+mL]w_{2L}[n]$, and $w_{2L}[n]$ is the rectangular window of length $2L$. We further assume that the whole frame of length $2L$ covers Q cycles of the resampling delay curve in Fig. 13, where the period of each cycle is $2N$ and Q is a positive integer, and we have the relationship $L = QN$.

We know from (118) that the sub-sample delay for an expansion ratio $R > 1$ is given by $(R-1)n/R = \Delta Rn/R$. The sub-sample delay for a compression ratio $R' = 1/R$ can be similarly given by $(R'-1)n/R' = -\Delta Rn$. Therefore, we can formulate the resampling delay curves in Fig. 13 as

$$d_1(n) = \begin{cases} \frac{\Delta R}{R}(n - 2qN), & n \in [2qN, (2q+1)N - 1] \\ \Delta R[(2q+2)N - n], & n \in [(2q+1)N, (2q+2)N - 1] \end{cases} \tag{191}$$

$$d_2(n) = \begin{cases} \Delta R(2qN - n), & n \in [2qN, (2q+1)N - 1] \\ \frac{\Delta R}{R}[n - (2q+2)N], & n \in [(2q+1)N, (2q+2)N - 1], \end{cases} \tag{192}$$

where $d_p(n)$, $p = 1, 2$, is the resampling delay in the p^{th} channel and $q = 0, \ldots, Q-1$.

Using the time-shifting property of the discrete Fourier transform (DFT), the resampled signals in both channels are given by

$$\tilde{x}_{p,m}[n] = \mathrm{IDFT}\{U_m[l]e^{-j\frac{\pi}{L}ld_p(n)}\} = \frac{1}{2L}\sum_{l=-L}^{L-1} U_m[l]e^{j\frac{\pi}{L}l(n-d_p(n))}, \quad p = 1, 2, \tag{193}$$

and the DFT coefficients of the resampled signals are

$$\tilde{X}_{p,m}[k] = \mathrm{DFT}\{\tilde{x}_{p,m}[n]\} = \sum_{n=0}^{2L-1}\tilde{x}_{p,m}[n]e^{-j\frac{\pi}{L}kn} = \frac{1}{2L}\sum_{l=-L}^{L-1} U_m[l]\sum_{n=0}^{2L-1} e^{j\frac{\pi}{L}[l(n-d_p(n))-kn]}, \tag{194}$$

where $k = 0, \ldots, 2L-1$. The last term of (194) for channel one can be written as

$$\begin{aligned}
&\sum_{n=0}^{2L-1} e^{j\frac{\pi}{L}[l(n-d_1(n))-kn]} \\
&= \sum_{q=0}^{Q-1}\left(\sum_{n=2qN}^{(2q+1)N-1} e^{j\frac{\pi}{L}[(\frac{l}{R}-k)n+\frac{\Delta R}{R}2qNl]} + \sum_{n=(2q+1)N}^{(2q+2)N-1} e^{j\frac{\pi}{L}[(Rl-k)n-\Delta R(2q+2)Nl]}\right) \\
&= \left(\sum_{n=0}^{N-1} e^{j\frac{\pi}{L}(\frac{l}{R}-k)n}\right)\left(\sum_{q=0}^{Q-1} e^{j\frac{2\pi N}{L}(l-k)q}\right) \\
&\quad + \left(\sum_{n=0}^{N-1} e^{j\frac{\pi}{L}(Rl-k)n}\right)\left(\sum_{q=0}^{Q-1} e^{j\frac{2\pi N}{L}(l-k)q}\right)e^{j\frac{\pi N}{L}[(1-\Delta R)l-k]}.
\end{aligned} \tag{195}$$

The last equality in (195) can be simplified by using

$$\sum_{n=0}^{N-1} e^{j\frac{\pi}{L}(\frac{l}{R}-k)n} = \frac{\sin\big(\frac{\pi}{2Q}\big(\frac{l}{R}-k\big)\big)}{\sin\big(\frac{1}{N}\frac{\pi}{2Q}\big(\frac{l}{R}-k\big)\big)}e^{j\frac{N-1}{N}\frac{\pi}{2Q}(\frac{l}{R}-k)} = \phi_N\big(\tfrac{1}{2Q}\big(\tfrac{l}{R}-k\big)\big) \tag{196}$$

$$\sum_{n=0}^{N-1} e^{j\frac{\pi}{L}(Rl-k)n} = \phi_N\big(\tfrac{1}{2Q}(Rl-k)\big) \tag{197}$$

$$\sum_{q=0}^{Q-1} e^{j\frac{2\pi N}{L}(l-k)q} = \frac{\sin(\pi(l-k))}{\sin\big(\frac{1}{Q}\pi(l-k)\big)}e^{j\frac{Q-1}{Q}\pi(l-k)} = \phi_Q(l-k), \tag{198}$$

and we have

$$\begin{aligned}
\sum_{n=0}^{2L-1} e^{j\frac{\pi}{L}[l(n-d_1(n))-kn]} &= \phi_N\big(\tfrac{1}{2Q}\big(\tfrac{l}{R}-k\big)\big)\phi_Q(l-k) \\
&\quad + \phi_N\big(\tfrac{1}{2Q}(Rl-k)\big)\phi_Q(l-k)e^{j\frac{1}{Q}\pi[(1-\Delta R)l-k]}.
\end{aligned} \tag{199}$$

Similarly for channel two, we have

$$\sum_{n=0}^{2L-1} e^{j\frac{\pi}{L}[l(n-d_2(n))-kn]} = \phi_N\big(\tfrac{1}{2Q}(Rl-k)\big)\phi_Q(l-k) + \phi_N\big(\tfrac{1}{2Q}\big(\tfrac{l}{R}-k\big)\big)\phi_Q(l-k)e^{j\frac{1}{Q}\pi[(1+\frac{\Delta R}{R})l-k]}. \quad (200)$$

Using (199) and (200) for the last term of (194), we obtain (130).

APPENDIX B

CROSS-SPECTRAL DENSITY OF WHITE GAUSSIAN NOISE AFTER RESAMPLING: WITH FAR-END ROOM IMPULSE RESPONSE

Let $u[n]$ be a zero-mean white Gaussian noise (WGN) with variance σ_u^2 and $g_p[n]$, $p = 1, 2$, be the far-end room impulse response. The filtering operation at the far-end room is analyzed in the frequency domain using the circular convolution theorem of the discrete Fourier transform (DFT). Here we assume $K \geq 2L$, where L is the length of the near-end room impulse response and K is the length of the far-end room impulse response. The reason for this assumption is that for a usable filtered output of length $2L$ at the far-end (which is required to meet the overlap-save frequency-domain adaptive filter (FDAF) structure [102] for the near-end stereophonic acoustic echo cancellation (SAEC)), we enforce an input data length of at least $4L$. We can zero-pad the actual far-end room impulse response if it is smaller than $2L$. The far-end room impulse response can be expressed in the frequency domain as

$$G_p[s] = \sum_{n=0}^{2K-1} g_p[n] e^{-j\frac{\pi}{K}sn}, \quad p = 1, 2, \tag{201}$$

where $s = -K, \ldots, K-1$ and $g_p[n] = 0$, $n \geq K$. The discrete short-time Fourier transform (STFT) of the WGN is

$$U_m[s] = e^{-j\frac{\pi L}{K}sm} \sum_{n=0}^{2K-1} u_m[n] e^{-j\frac{\pi}{K}sn}, \tag{202}$$

where $u_m[n] = u[n + mL]w_{2K}[n]$ and $w_{2K}[n]$ is the rectangular window of length $2K$. Note that we shift the filtering operation of the noise block by L to produce a 50% overlap for a usable filtered output of length $2L$ as the reference signal.

Since the multiplication in the DFT domain is equal to the circular convolution in the time domain, the usable filtered output is given by

$$(g_p * u_m)[n] = \frac{1}{2K} \sum_{s=-K}^{K-1} G_p[s] U_m[s] e^{j\frac{\pi}{K}sn}, \tag{203}$$

where $n = 2(K-L), \ldots, 2K-1$ and only the last $2L$ samples are the usable filtered output. By rearranging the sample indices, the reference signals in the time domain can be written as

$$x_{p,m}[n] = (g_p * u_m)[n + 2(K-L)] = \frac{1}{2K} \sum_{s=-K}^{K-1} G_p[s] U_m[s] e^{-j\frac{2\pi L}{K}s} e^{j\frac{\pi}{K}sn}, \tag{204}$$

where $n = 0, \ldots, 2L-1$. Therefore, the DFT coefficients of this block of filtered noise can be written as

$$X_{p,m}[l] = \sum_{n=0}^{2L-1} x_{p,m}[n] e^{-j\frac{\pi}{L}ln} = \frac{1}{2K} \sum_{s=-K}^{K-1} G_p[s] U_m[s] e^{-j\frac{2\pi L}{K}s} \phi_{2L}\big(\tfrac{L}{K}s - l\big), \tag{205}$$

where $l = -L, \ldots, L-1$. Using (130), the resampled reference signals are given by

$$\begin{aligned} \tilde{X}_{p,m}[k] &= \frac{1}{2L} \sum_{l=-L}^{L-1} X_{p,m}[l] \Phi_p(k,l) \\ &= \frac{1}{4KL} \sum_{l} \Phi_p(k,l) \sum_{s} G_p[s] U_m[s] e^{-j\frac{2\pi L}{K}s} \phi_{2L}\big(\tfrac{L}{K}s - l\big), \end{aligned} \tag{206}$$

where $k = 0, \ldots, 2L-1$. The cross-spectral density between the resampled reference signals is given by

$$\begin{aligned} &\mathrm{E}\{\tilde{X}_{1,m}[k] \tilde{X}^*_{2,m}[k]\} \\ &= \frac{1}{16K^2L^2} \sum_{l_1,l_2} \Phi_1(k,l_1) \Phi_2^*(k,l_2) \\ &\quad \cdot \sum_{s_1,s_2} G_1[s_1] G_2^*[s_2] e^{j\frac{2\pi L}{K}(s_2 - s_1)} \phi_{2L}\big(\tfrac{L}{K}s_1 - l_1\big) \phi^*_{2L}\big(\tfrac{L}{K}s_2 - l_2\big) \mathrm{E}\{U_m[s_1] U_m^*[s_2]\} \\ &= \frac{\sigma_u^2}{8KL^2} \sum_{s=-K}^{K-1} G_1[s] G_2^*[s] \left(\sum_{l=-L}^{L-1} \Phi_1(k,l) \phi_{2L}\big(\tfrac{L}{K}s - l\big) \right) \left(\sum_{l=-L}^{L-1} \Phi_2(k,l) \phi_{2L}\big(\tfrac{L}{K}s - l\big) \right)^*, \end{aligned} \tag{207}$$

where the last equality is obtained by using (125). Substituting the last two terms of (207) by (141), we obtain (140).

REFERENCES

[1] Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I., "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1–2, pp. 5–43, 2003.

[2] Benesty, J. and Duhamel, P., "Fast constant modulus adaptive algorithm," *Radar and Signal Processing, IEE Proceedings F*, vol. 138, no. 4, pp. 379–387, Aug. 1991.

[3] Benesty, J. and Duhamel, P., "A fast exact least mean square adaptive algorithm," *Signal Processing, IEEE Transactions on*, vol. 40, no. 12, pp. 2904–2920, Dec. 1992.

[4] Benesty, J. and Gansler, T., "A robust fast recursive least squares adaptive algorithm," *Acoustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on*, vol. 6, pp. 3785–3788, May 2001.

[5] Benesty, J., Gansler, T., Morgan, D. R., Sondhi, M. M., and Gay, S. L., *Advances in Network and Acoustic Echo Cancellation*. Berlin, Germany: Springer-Verlag, May 2001.

[6] Benesty, J., Morgan, D. R., and Cho, J. H., "A new class of doubletalk detectors based on cross-correlation," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 2, pp. 168–172, Mar. 2000.

[7] Benesty, J., Morgan, D. R., and Sondhi, M. M., "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[8] Benesty, J., Paleologu, C., and Ciochina, S., "On regularization in adaptive filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1734–1742, Aug 2011.

[9] Birkenes, O., "A phase robust spectral magnitude estimator for acoustic echo suppression," *Acoustic Echo and Noise Control (IWAENC), 2010 International Workshop on*, Aug. 2010.

[10] Boyd, S. and Vandenberghe, L., *Convex Optimization*. New York, NY, USA: Cambridge University Press, Mar. 2004.

[11] Brady, P. T., "A statistical analysis of on-off patterns in 16 conversations," *Bell System Technical Journal, The*, vol. 47, no. 1, pp. 73–91, Jan. 1968.

[12] Brady, P. T., "A model for generating on-off speech patterns in two-way conversation," *Bell System Technical Journal, The*, vol. 48, no. 7, pp. 2445–2472, Sept. 1969.

[13] Brandwood, D. H., "A complex gradient operator and its application in adaptive array theory," *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[14] Buchner, H., Benesty, J., and Gansler, T., "An outlier-robust extended multidelay filter with application to acoustic echo cancellation," *Acoustic Echo and Noise Control (IWAENC), 2003 International Workshop on*, pp. 19–22, Sept. 2003.

[15] Butz, M. V., Sastry, K., and Goldberg, D. E., "Tournament selection: Stable fitness pressure in XCS," *Genetic and Evolutionary Computation COnference (GECCO)*, vol. 2724, pp. 1857–1869, July 2003.

[16] Carter, G. C., "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, no. 2, pp. 236–255, Feb. 1987.

[17] Chhetri, A. S., Surendran, A. C., Stokes, J. W., and Platt, J. C., "Regression-based residual acoustic echo suppression," *Acoustic Echo and Noise Control (IWAENC), 2005 International Workshop on*, pp. 201–204, Sept. 2005.

[18] Christensen, M. G. and Jensen, S. H., "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 99–109, Jan. 2006.

[19] Cohen, I., "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[20] Cohen, I. and Berdugo, B., "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *Signal Processing Letters, IEEE*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[21] Cooke, M., Green, P., Josifovski, L., and Vizinho, A., "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, June 2001.

[22] Deb, K., "An efficient constraint handling method for genetic algorithms," *Computer Methods in Applied Mechanics and Engineering*, vol. 186, no. 2–4, pp. 311–338, June 2000.

[23] Deshmukh, O. D., "Embedded automatic speech recognition and text-to-speech synthesis," in *Speech in Mobile and Pervasive Environments* (Rajput, N. and Nanavati, A. A., eds.), Chichester, UK.: John Wiley & Sons, Ltd, Jan. 2012.

[24] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*. Wiley-Interscience, 2nd ed., Nov. 2000.

[25] Duttweiler, D. L., "A twelve-channel digital echo canceler," *Communications, IEEE Transactions on*, vol. 26, no. 5, pp. 647–653, May 1978.

[26] Enzner, G., Martin, R., and Vary, P., "On spectral estimation of residual echo in hands-free telephony," *Acoustic Echo and Noise Control (IWAENC), 2001 International Workshop on*, Sept. 2001.

[27] Enzner, G., Martin, R., and Vary, P., "Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 103–114, Mar. 2002.

[28] Enzner, G., Martin, R., and Vary, P., "Unbiased residual echo power estimation for hands-free telephony," *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, pp. 1893–1896, May 2002.

[29] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[30] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[31] Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S., Voss, C., and Zeller, F., "Automatic human utility evaluation of ASR systems: does WER really predict performance?," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, pp. 3463–3467, Aug. 2013.

[32] Gansler, T., "A double-talk resistant subband echo canceller," *Signal Processing*, vol. 65, no. 1, pp. 89–101, Feb. 1998.

[33] Gansler, T. and Benesty, J., "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," *International Journal of Adaptive Control and Signal Processing*, vol. 14, no. 6, pp. 565–586, Sept. 2000.

[34] Gansler, T. and Benesty, J., "New insights into the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 257–267, July 2002.

[35] Gansler, T. and Benesty, J., "The fast normalized cross-correlation double-talk detector," *Signal Processing*, vol. 86, no. 6, pp. 1124–1139, June 2006.

[36] Gansler, T., Gay, S. L., Sondhi, M. M., and Benesty, J., "Double-talk robust fast converging algorithms for network echo cancellation," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 656–663, Nov. 2000.

[37] Gansler, T., Hansson, M., Ivarsson, C. J., and Salomonsson, G., "A double-talk detector based on coherence," *Communications, IEEE Transactions on*, vol. 44, no. 11, pp. 1421–1427, Nov. 1996.

[38] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[39] Gerkmann, T. and Hendriks, R. C., "Noise power estimation based on the probability of speech presence," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 145–148, Oct. 2011.

[40] Gerkmann, T. and Hendriks, R. C., "Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 105–108, Mar. 2012.

[41] Gerkmann, T. and Hendriks, R. C., "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[42] Giacobello, D., Atkins, J., Wung, J., and Prabhu, R., "Results on automated tuning of a voice quality enhancement system using objective quality measures," in *Audio Engineering Society Convention 135*, Oct. 2013.

[43] Giacobello, D., Wung, J., Pichevar, R., and Atkins, J., "A computationally constrained optimization framework for implementation and tuning of speech enhancement systems," in *Acoustic Signal Enhancement (IWAENC), 2014 International Workshop on*, pp. 159–163, Sept. 2014.

[44] Giacobello, D., Wung, J., Pichevar, R., and Atkins, J., "Tuning methodology for speech enhancement algorithms using a simulated conversational database and perceptual objective measures," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, pp. 62–66, May 2014.

[45] Goetze, S., Kallinger, M., and Kammeyer, K.-D., "Residual echo power spectral density estimation based on an optimal smoothed misalignment for acoustic echo cancelation," *Acoustic Echo and Noise Control (IWAENC), 2005 International Workshop on*, pp. 209–212, Sept. 2005.

[46] Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1st ed., Jan. 1989.

[47] Gustafsson, S. and Martin, R., "Combined acoustic echo control and noise reduction based on residual echo estimation," *Acoustic Echo and Noise Control (IWAENC), 1997 International Workshop on*, Sept. 1997.

[48] Gustafsson, S., Martin, R., Jax, P., and Vary, P., "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 245–256, July 2002.

[49] Gustafsson, S., Martin, R., and Vary, P., "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21–32, Jan. 1998.

[50] Hammer, F., Reichl, P., and Raake, A., "Elements of interactivity in telephone conversations," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea*, pp. 1741–1744, Oct. 2004.

[51] Hansler, E. and Schmidt, G. U., "Hands-free telephones – joint control of echo cancellation and postfiltering," *Signal Processing*, vol. 80, no. 11, pp. 2295–2305, Nov. 2000.

[52] Hartmann, W., Narayanan, A., Fosler-Lussier, E., and Wang, D., "A direct masking approach to robust ASR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 1993–2005, Oct. 2013.

[53] Hendriks, R. C., Gerkmann, T., and Jensen, J., "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, Jan. 2013.

[54] Hendriks, R. C., Heusdens, R., and Jensen, J., "MMSE based noise PSD tracking with low complexity," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4266–4269, Mar. 2010.

[55] Herre, J., Buchner, H., and Kellermann, W., "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*, vol. 1, pp. 17–20, Apr. 2007.

[56] Hines, A., Skoglund, J., Kokaram, A., and Harte, N., "ViSQOL: The virtual speech quality objective listener," in *Acoustic Signal Enhancement (IWAENC), 2012 International Workshop on*, Sept. 2012.

[57] Hines, A., Skoglund, J., Kokaram, A., and Harte, N., "Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 3697–3701, May 2013.

[58] Hirano, A. and Sugiyama, A., "A noise-robust stochastic gradient algorithm with an adaptive step-size suitable for mobile hands-free telephones," in *Acoustics, Speech, and Signal Processing (ICASSP), 1995 International Conference on*, vol. 2, pp. 1392–1395, May 1995.

[59] Hu, Y. and Loizou, P. C., "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, Aug. 2007.

[60] Hu, Y. and Loizou, P. C., "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[61] Huber, P. J. and Ronchetti, E. M., *Robust Statistics*. Wiley, 2nd ed., Jan. 2009.

[62] Iqbal, M. A., Stokes, J. W., and Grant, S. L., "Normalized double-talk detection based on microphone and AEC error cross-correlation," in *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 360–363, July 2007.

[63] ISO/IEC 11172-3:1993, "Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — Part 3: Audio," International Organization for Standardization, Geneva, Switzerland, 1993.

[64] ITU-R BS.1534-2, "Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunication Union Radiocommunication Sector, Geneva, Switzerland, June 2014.

[65] ITU-T G.167, "Acoustic echo controllers," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, Mar. 1993.

[66] ITU-T P Series, "Telephone transmission quality, telephone installations, local line networks," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland.

[67] ITU-T P.56, "Objective measurement of active speech level," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, Dec. 2011.

[68] ITU-T P.59, "Artificial conversational speech," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, Mar. 1993.

[69] ITU-T P.800, "Methods for subjective determination of transmission quality," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, Aug. 1996.

[70] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, Nov. 2003.

[71] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," international telecommunication union telecommunication standardization sector, Geneva, Switzerland, Feb. 2001.

[72] ITU-T P.863, "Perceptual objective listening quality assessment," International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland, Sept. 2014.

[73] Juang, B.-H., "Speech recognition in adverse environments," *Computer Speech & Language*, vol. 5, no. 3, pp. 275–294, July 1991.

[74] Kalinli, O., Seltzer, M. L., Droppo, J., and Acero, A., "Noise adaptive training for robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1889–1901, Nov. 2010.

[75] Khong, A. W. H., Benesty, J., and Naylor, P. A., "Stereophonic acoustic echo cancellation: analysis of the misalignment in the frequency domain," *Signal Processing Letters, IEEE*, vol. 13, no. 1, pp. 33–36, Jan. 2006.

[76] Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., and Maas, R., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct. 2013.

[77] Kronlid, F., "Turn taking for artificial conversational agents," in *Cooperative Information Agents X* (Klusch, M., Rovatsos, M., and Payne, T. R., eds.), vol. 4149 of *Lecture Notes in Computer Science*, pp. 81–95, Springer Berlin Heidelberg, 2006.

[78] Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R., "An overview of noise-robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[79] Liptak, B. G., *Instrument Engineers' Handbook, Volume 2: Process Control and Optimization*. CRC press, 4th ed., Sept. 2005.

[80] Loizou, P. C., *Speech Enhancement: Theory and Practice*. CRC press, 2nd ed., Feb. 2013.

[81] Lu, X. and Champagne, B., "A centralized acoustic echo canceller exploiting masking properties of the human ear," in *Acoustics, Speech, and Signal Processing (ICASSP), 2003 IEEE International Conference on*, vol. 5, pp. 377–380, Apr. 2003.

[82] Mader, A., Puder, H., and Schmidt, G. U., "Step-size control for acoustic echo cancellation filters – an overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, Sept. 2000.

[83] Mansour, D. and Gray, A. H., "Unconstrained frequency-domain adaptive filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 30, no. 5, pp. 726–734, Oct. 1982.

[84] Markovica, D., Antonacci, F., Sarti, A., and Tubaro, S., "Estimation of room dimensions from a single impulse response," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct. 2013.

[85] Martin, R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, July 2001.

[86] Martin, R., "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, June 2006.

[87] Meyer-Baese, U., *Digital Signal Processing with Field Programmable Gate Arrays*. Springer, 3rd ed., Apr. 2007.

[88] Moller, S., Chan, W.-Y., Cote, N., Falk, T. H., Raake, A., and Waltermann, M., "Speech quality estimation: Models and trends," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, Nov. 2011.

[89] Morgan, D. R., Hall, J. L., and Benesty, J., "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 6, pp. 686–696, Sept. 2001.

[90] Moulines, E., Ait Amrane, O., and Grenier, Y., "The generalized multidelay adaptive filter: structure and convergence analysis," *Signal Processing, IEEE Transactions on*, vol. 43, no. 1, pp. 14–28, Jan. 1995.

[91] Nitsch, B. H., "A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain," *Signal Processing*, vol. 80, no. 9, pp. 1733–1745, Sept. 2000.

[92] Oppenheim, A. V. and Schafer, R. W., *Discrete-Time Signal Processing*. Prentice Hall, 3rd ed., Aug. 2009.

[93] Puschel, M., Moura, J. M. F., Johnson, J. R., Padua, D., Veloso, M. M., Singer, B. W., Xiong, J., Franchetti, F., Gacic, A., Voronenko, Y., Chen, K., Johnson, R. W., and Rizzolo, N., "SPIRAL: Code generation for DSP transforms," *Proceedings of the IEEE*, vol. 93, no. 2, pp. 232–275, Feb. 2005.

[94] Rabiner, L. R. and Juang, B.-H., *Fundamentals of Speech Recognition*. Prentice Hall, 1st ed., Apr. 1993.

[95] Raj, B., Seltzer, M. L., and Stern, R. M., "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, Sept. 2004.

[96] Raux, A. and Eskenazi, M., "A finite-state turn-taking model for spoken dialog systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Boulder, CO, USA), pp. 629–637, June 2009.

[97] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P., "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing (ICASSP), 2001 IEEE International Conference on*, vol. 2, pp. 749–752, May 2001.

[98] Robledo-Arnuncio, E., Wada, T. S., and Juang, B.-H., "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2007 IEEE Workshop on*, pp. 34–37, Oct. 2007.

[99] Sayed, A. H., *Fundamentals of Adaptive Filtering*. Wiley-IEEE Press, 1st ed., June 2003.

[100] Sayed, A. H., *Adaptive Filters*. Wiley-IEEE Press, 1st ed., Apr. 2008.

[101] Schroeder, M. R., "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[102] Shynk, J. J., "Frequency-domain and multirate adaptive filtering," *Signal Processing Magazine, IEEE*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[103] Smith, S. W., *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Pub, 1st ed., 1997.

[104] Sondhi, M. M., Morgan, D. R., and Hall, J. L., "Stereophonic acoustic echo cancellation – an overview of the fundamental problem," *Signal Processing Letters, IEEE*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[105] Soo, J.-S. and Pang, K. K., "Multidelay block frequency domain adaptive filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 2, pp. 373–376, Feb. 1990.

[106] Sugiyama, A., Joncour, Y., and Hirano, A., "A stereo echo canceler with correct echo-path identification based on an input-sliding technique," *Signal Processing, IEEE Transactions on*, vol. 49, no. 11, pp. 2577–2587, Nov. 2001.

[107] Sugiyama, A., Mizuno, Y., Hirano, A., and Nakayama, K., "A stereo echo canceller with simultaneous input-sliding and sliding-period control," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 325–328, Mar. 2010.

[108] Tashev, I. J., *Sound Capture and Processing: Practical Approaches*. Wiley, 1st ed., Aug. 2009.

[109] Tashev, I. J., "Coherence based double talk detector with adaptive threshold," *20th International Scientific Conference Electronics ET 2011*, Sept. 2011.

[110] Tashev, I. J., "Coherence based double talk detector with soft decision," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 165–168, Mar. 2012.

[111] Tashev, I. J., Lovitt, A., and Acero, A., "Unified framework for single channel speech enhancement," in *Communications, Computers and Signal Processing, 2009 IEEE Pacific Rim Conference on*, pp. 883–888, Aug. 2009.

[112] Tashev, I. J. and Slaney, M., "Data driven suppression rule for speech enhancement," in *Proceedings of the Information Theory and Applications Workshop*, pp. 1–6, Feb. 2013.

[113] Turbin, V., Gilloire, A., Scalart, P., and Beaugeant, C., "Using psychoacoustic criteria in acoustic echo cancellation algorithms," *Acoustic Echo and Noise Control (IWAENC), 1997 International Workshop on*, pp. 53–56, Sept. 1997.

[114] Wada, T. S. and Juang, B.-H., "Enhancement of residual echo for improved acoustic echo cancellation," in *15th European Signal Processing Conference (EUSIPCO 2007)*, pp. 1620–1624, Sept. 2007.

[115] Wada, T. S. and Juang, B.-H., "Enhancement of residual echo for improved frequency-domain acoustic echo cancellation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2007 IEEE Workshop on*, pp. 175–178, Oct. 2007.

[116] Wada, T. S. and Juang, B.-H., "Towards robust acoustic echo cancellation during double-talk and near-end background noise via enhancement of residual echo," in *Acoustics, Speech and Signal Processing (ICASSP), 2008 IEEE International Conference on*, pp. 253–256, Mar. 2008.

[117] Wada, T. S. and Juang, B.-H., "Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2009 IEEE Workshop on*, pp. 205–208, Oct. 2009.

[118] Wada, T. S. and Juang, B.-H., "Multi-channel acoustic echo cancellation based on residual echo enhancement with effective channel decorrelation via resampling," in *Acoustic Echo and Noise Control (IWAENC), 2010 International Workshop on*, Aug. 2010.

[119] Wada, T. S. and Juang, B.-H., "Enhancement of residual echo for robust acoustic echo cancellation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 175–189, Jan. 2012.

[120] Wada, T. S., Wung, J., and Juang, B.-H., "Decorrelation by resampling in frequency domain for multi-channel acoustic echo cancellation based on residual echo enhancement," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 289–292, Oct. 2011.

[121] Watanabe, S. and Le Roux, J., "Black box optimization for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 3256–3260, May 2014.

[122] Welch, P. D., "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, no. 2, pp. 70–73, June 1967.

[123] Wung, J., Wada, T. S., and Juang, B.-H., "Inter-channel decorrelation by resampling via time-domain interpolation filters derived from the time-shifting property," in *Acoustic Signal Enhancement (IWAENC), 2012 International Workshop on*, Sept. 2012.

[124] Wung, J., Wada, T. S., and Juang, B.-H., "Inter-channel decorrelation by sub-band resampling in frequency domain," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 29–32, Mar. 2012.

[125] Wung, J., Wada, T. S., and Juang, B.-H., "On the performance of the robust acoustic echo cancellation system with decorrelation by sub-band resampling," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 635–638, May 2013.

[126] Wung, J., Wada, T. S., Juang, B.-H., Lee, B., Kalker, T., and Schafer, R. W., "A system approach to residual echo suppression in robust hands-free teleconferencing," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 445–448, May 2011.

[127] Wung, J., Wada, T. S., Juang, B.-H., Lee, B., Trott, M. D., and Schafer, R. W., "A system approach to acoustic echo cancellation in robust hands-free teleconferencing," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 101–104, Oct. 2011.

[128] Wung, J., Wada, T. S., Souden, M., and Juang, B.-H., "On the misalignment of stereophonic acoustic echo cancellation with decorrelation by resampling," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct. 2013.

[129] Wung, J., Wada, T. S., Souden, M., and Juang, B.-H., "Inter-channel decorrelation by sub-band resampling for multi-channel acoustic echo cancellation," *Signal Processing, IEEE Transactions on*, vol. 62, no. 8, pp. 2127–2142, Apr. 2014.

[130] Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., *The HTK Book Version 3.4*. Cambridge University Press, Dec. 2006.